RESEARCH & INNOVATION DIVISION IJIST

Comparative Study of Machine Learning Algorithms for Sentiment Analysis in Multimodal Medical Data

Hafiz Muhammad Bilal[1], Muhammad Asif[2], Muhammad Azam Zia[2]

[1]University of Agriculture Faisalabad

[2]Information Technology University

***Correspondence:** Muhammad Asif, Muhammad Azam Zia

muhammadasif.agri@hotmail.com, mazamzia@uaf.edu.pk

Citation | Bilal. H. M, Asif. M, Zia. M. A, "Comparative Study of Machine Learning Algorithms for Sentiment Analysis in Multimodal Medical Data", IJIST, Special Issue. pp 304-317, 2025

Received March 01, 2025 **Revised** | March 13, 2025 **Accepted** | March 19, 2025 **Published** | March 22, 2025.

Sentiment analysis, a part of data mining, uses Natural Language Processing (NLP) to understand how people feel about certain topics or individuals. It focuses on the context and polarity of information, measuring public opinions from unstructured sources like social networks and healthcare websites. By extracting useful insights from this unstructured data, healthcare professionals can improve patient care, make accurate diagnoses, and provide personalized treatments. Machine learning (ML) plays a key role in this process. ML techniques like logistic regression, decision trees, and Naive Bayes have proven effective in tasks such as sentiment analysis and named entity recognition in medical data. The goal of ML is to create algorithms that enhance data processing and decision-making by identifying patterns that might be overlooked by humans. In this study, we compare the performance of three common ML models—(a) Logistic Regression, (b) Decision Tree, and (c) Naive Bayes—for sentiment analysis on medical image captions. The Radiology Objects in Context (ROCO) multimodal image and caption dataset was used for this NLP task. Caption pre-processing is done using filtering methods to improve text quality, followed by sentiment classification using pre-trained ML models. This comparison sheds light on the effectiveness of these algorithms in performing sentiment analysis in clinical settings.

Keywords: Sentiment Analysis; Machine Learning, Natural Language Processing, Confusion Matrix.

Introduction:

Electronic Health Records (EHRs) are widely used worldwide and provide valuable resources for research, improving healthcare quality and population management. As the number of EHRs continues to grow rapidly, effective data analysis methods become increasingly important. Narrative reports, which are the core components of EHR systems, provide detailed information about patients' conditions, reasons for treatment, and doctor-patient interactions [1]. These details are often too complex for structured tables, requiring advanced techniques for analysis.

Natural Language Processing (NLP) has proven effective in processing the text data extracted from EHR systems. It helps extract key information, such as medications and diagnoses, typically in the form of single words or short phrases. However, NLP still faces challenges in extracting more complex information, such as understanding the relationship between illnesses and symptoms, resolving ambiguity in medical terminology, or analyzing emotional tones over multiple sentences [2]. This is where machine learning (ML), a subfield of AI, becomes crucial.

ML uses data-driven algorithms to improve computer performance in tasks like decision-making and pattern recognition. ML models can identify trends and patterns that humans might miss when analyzing historical data. In healthcare, ML has facilitated communication between medical professionals and computer scientists, especially through data mining [3]. Data mining extracts valuable insights from large datasets, reducing the need for costly and invasive medical procedures while improving efficiency and cost-effectiveness [4]. For instance, data mining can help detect high-risk patients and identify key factors associated with positive or negative health outcomes, without relying solely on invasive procedures like X-rays, blood tests, or angiograms.

In healthcare, there is a need to perform sentiment analysis on multimodal clinical data to gain meaningful insights from text-image pairs. This can help determine the best sentiment analysis techniques, improve interpretability, and assist decision-making. It is valuable to compare and contrast different machine-learning models for these tasks.

The use of data mining and machine learning in disease identification and prediction has increased in recent years. Their complexity and application have significantly reduced medical errors and improved diagnostic accuracy. Among the most widely used models for these tasks are logistic regression, decision trees, and Naïve Bayes classifiers.

Logistic Regression: Logistic Regression (LR) is a statistical model used for categorical outcomes, such as binary classifications (e.g., yes/no, true/false). It estimates the probability of an event occurring based on a linear combination of explanatory variables. Its flexibility comes from the minimal assumptions it makes about these variables, making it a useful tool for healthcare predictions [5].

Naïve Bayes: The Naïve Bayes (NB) classifier uses Bayes' theorem to calculate the probability that data belongs to a specific category. It assumes that the features are independent, which simplifies calculations and allows for efficient predictions. Despite this "naïve" assumption, it often provides strong performance in real-world applications [6].

Decision Trees:

Decision Trees (DT) are commonly used for both classification and regression tasks. They divide data into smaller subsets based on feature values, with decisions represented as nodes and outcomes as branches. This clear structure makes Decision Trees particularly useful in healthcare, where understanding the reasoning behind predictions is crucial [7]. In this study, we compare the performance of three machine learning models—Logistic Regression, Decision Trees, and Naïve Bayes—in analyzing text data from Electronic Health Records (EHRs). Our goal is to identify key diagnostic factors and evaluate the accuracy of these models in predicting the need for medical interventions [8]. The results aim to enhance healthcare practices by leveraging ML's ability to support informed and effective decision-making.

The Unified Modelling Language (UML) system facilitates interoperability among different medical terminologies by linking concepts from various databases, providing an integrated view of medical information. The dataset consists of 8,179 image captions, with 11,154 unique tokens (features) representing the vocabulary. The longest caption contains 133 words, with an average length of about 99.64 words. This dense dataset includes both images and text.

This paper explores Neural Image Captioning (NIC) and its applications in radiology, specifically focusing on sentiment analysis of multimodal medical text data. While previous studies have focused on generating textual descriptions from medical images, this research examines the emotional and subjective aspects of medical texts, such as medical reports, captions, and EHRs. The study adopts a comparative and experimental approach, evaluating several machine learning algorithms for sentiment analysis in the medical domain. By addressing sentiment analysis, a relatively under-researched area in medical AI, this work fills a gap in understanding the emotional and subjective dimensions of medical texts. By using multimodal inputs and comparing traditional and advanced machine learning techniques, the study offers valuable insights and benchmarks for improving sentiment analysis methods in clinical applications. This unique focus not only complements existing research but also provides new perspectives and practical implications for medical AI.

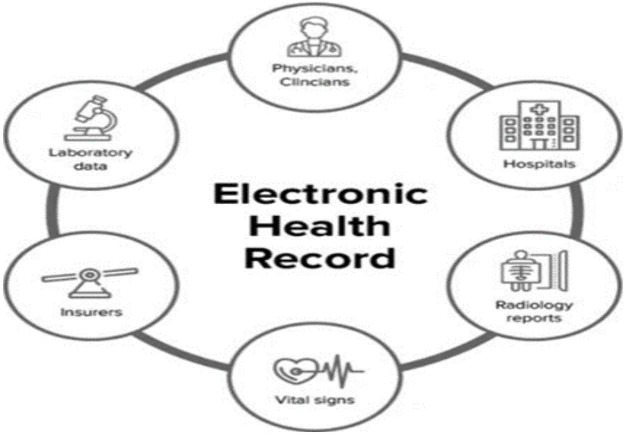

Figure 1. EHR Health Care Application

Objectives of the Study:

The primary objectives of this study are:

1. To use the Radiology Objects in Context (ROCO) dataset to explore multimodal image data and corresponding text descriptions. The goal is to process and standardize the image captions using Natural Language Processing (NLP) tasks to improve data quality for subsequent sentiment analysis.
2. To perform sentiment analysis on the captions of medical images, categorizing them into three classes: positive, negative, and neutral. The study uses the VADER sentiment analysis model to label the data and assess sentiment based on the compound sentiment score derived from individual word sentiment scores.
3. To apply the TF-IDF technique to convert textual data into numerical representations, aiding in the identification of significant terms and enhancing machine learning model performance.
4. To evaluate the performance of different machine learning models (Logistic Regression, Decision Tree, and Naïve Bayes) in classifying sentiment in medical image captions. This includes testing each model’s accuracy, precision, recall, and F1 score, focusing on comparing their ability to classify the three sentiment categories.

5. To use confusion matrices and standard evaluation metrics such as accuracy, precision, recall, and F1 score to measure the efficiency of the models and ensure proper sentiment classification in medical image captions.

Material and Methods:

This cross-sectional study used the Radiology Objects in Context (ROCO) dataset, which contains a large-scale multimodal image collection. The images in this dataset are sourced from the PubMed "Central Open Access FTP" mirror, making them publicly accessible and suitable for broad research applications. The images are categorized into two types: non-compound and radiological, based on their identification. In addition to the images, the ROCO dataset includes rich metadata, such as Unified Modelling Language (UML) semantic types, UMLS Concepts Unique Identifiers (CUIs), and image caption keywords [9]. This metadata enhances the understanding of image captions and their relevance in medical text, facilitating the integration of structured medical entities. The approach allows for the comparison and evaluation of machine learning models for sentiment analysis of medical captions, improving the comprehension of health data.

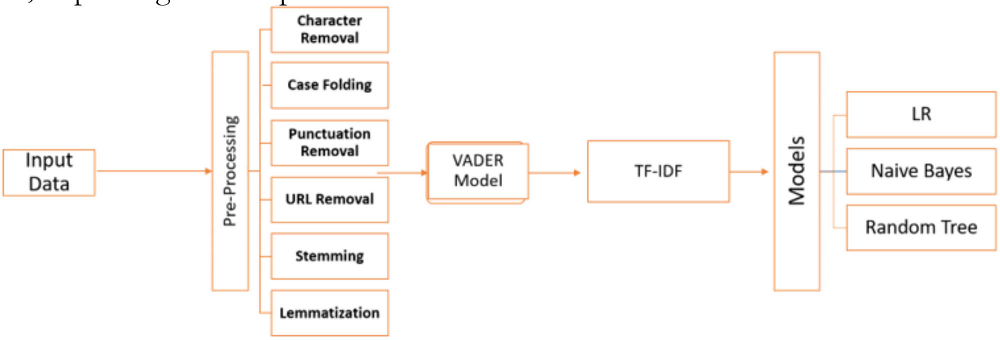

Figure 2. Methodology Flow Diagram

NLP Tasks:

The captions of the images in the dataset are pre-processed through a series of Natural Language Processing (NLP) tasks to standardize the data and remove textual noise, preparing it for further analysis. The dataset consists of two main columns:

- **ID Column**: This uniquely identifies each record in the dataset.
- **Description Column**: This contains the caption or text description for each image.

The Description Column is extracted and further processed for pre-processing [10].

The primary goal of the pre-processing strategies was to standardize the text and remove unnecessary or irrelevant elements, referred to as "textual noise." The key pre-processing steps included:

- **Character Removal**: All characters except letters, numbers, and spaces were removed (e.g., "coronal view" remains as "coronal view").
- **Case Folding** [11]: The text was converted to lowercase to ensure uniformity (e.g., "Axial MRI" was converted to "axial MRI").

Punctuation Removal:

Punctuation marks were removed (e.g., "Damus–Kaye–Stansel" became "damuskayestansel").

URL Removal: Any URLs within the captions were eliminated. Additionally, stop words (such as "is," "the," and "and") that do not contribute significantly to the meaning of sentiment analysis were removed from the data.

To further improve the data's quality for analysis, two key text normalization techniques were applied:

- **Stemming**: This method simplified words to their most basic forms (e.g., "running" became "run").

- **Lemmatization**: This technique was applied to ensure that words were reduced to their fundamental or dictionary form (e.g., "better" became "good") [12].

VADER Model for Data Labelling:

The valence-aware dictionary and Sentiment Reasoner (VADER) model was employed for the initial labeling of the text in terms of sentiment categories (positive, negative, or neutral). VADER is a rule-based algorithm that assigns sentiment scores to text using a pre-built sentiment lexicon [13]. The model evaluates sentiment on a scale from -1 to 1, where:

- **-1** indicates a strong negative sentiment.
- **0** indicates a neutral sentiment.
- **1** indicates a strong positive sentiment.

The sentiment score (S) is calculated using the following equation:

S=∑i=1n(wi·si)S = \sum_{i=1}^{n} (w_i \cdot s_i)S=i=1∑n(wi·si)

Where:

- wiw_iwi is the weight of the i-th word.
- sis_isi is the sentiment score of the i-th word in the lexicon.
- nnn is the number of words in the text.

$$\frac{\sum \text{Sentiment Score of Individual Words}}{\sqrt{\sum \text{Squared Sentiment Score of Individual Words}}}$$

The sentiment score for each word is assigned based on a sentiment lexicon. The sum of the sentiment scores gives an aggregated score for the entire text.

- The sentiment is considered positive if the compound score is greater than 0.05.
- The sentiment is considered negative if the compound score is lower than -0.05.
- The sentiment is considered neutral if the compound score falls within the range of -0.05 to 0.05.

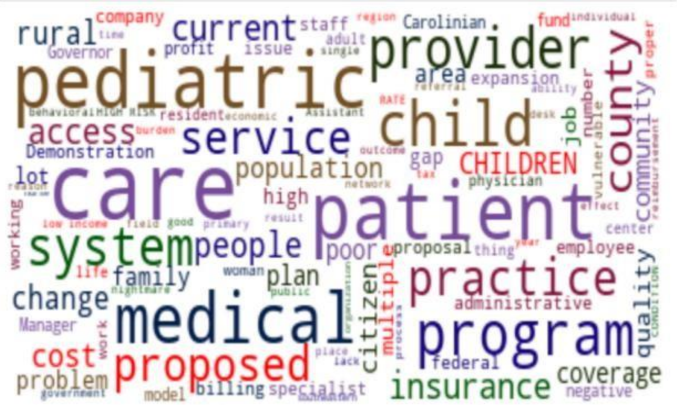

Figure 3. Word Cloud Sentiments

VADER enhances sentiment analysis by considering context at the word level using booster and negation words. For example, in the phrase "very happy," the word "very" acts as a booster to amplify the positive sentiment. Similarly, negation words such as "not" in "not happy" flip the sentiment polarity. This ability to modify sentiment based on contextual modifiers allows VADER to capture a more nuanced sense of sentiment in text. In this research, sentiment analysis is conducted in three classes: positive, negative, and neutral, with the sentiment classification determined by the compound score. The compound score is an aggregated score calculated by adding up the weighted sentiment scores of words in the text. This score is then mapped to one of the three sentiment categories, as shown in Table 1.

Vectorization:

Text data is modeled using the TF-IDF (Term Frequency-Inverse Document Frequency) technique, which is widely used to convert textual features into numerical ones. This method helps assess the importance of a word in a document relative to the overall corpus by scaling the frequency of the word within the document against how often it occurs across all texts [14].

- **Term Frequency (TF)**: The occurrence of a word ii in a given document dd, which indicates how often a specific word appears in a document.
- **Inverse Document Frequency (IDF)**: Measures the word's relevance across the entire corpus by calculating the logarithm of the ratio of the total number of documents in the corpus NN to the number of documents NiN_i that contain the word ii. This is used to calculate the word's significance.

The mathematical formulation for TF-IDF is:

TFIDF(i,d)=TF(i,d)×log⁡(NNi)TFIDF(i, d) = TF(i, d) \times \log \left(\frac{N}{N_i}\right)

Where:

- f(i,d)f(i, d) is the frequency of the word ii in document dd.
- NN is the total number of documents in the corpus.
- NiN_i is the number of documents that include the word ii.

By applying the TF-IDF method, the term frequency is multiplied by the inverse document frequency to highlight words that are specific to a document and reduce the influence of common words across the corpus. This approach ensures that words that appear frequently in a document but are rare across the corpus are given higher importance in the feature matrix. The TF-IDF method enables the system to emphasize unique words in each document while reducing the weight of generic terms that appear frequently across the entire corpus.

Table 1. Classification of Medical Imaging Based on Compound Score

	Compound Score	Class
Abdomen computed tomography like cholecystocutan fistula track	0.3612	Positive
axial mri coron view coron plain computed tomography images show multiple large tumor masses edge enhanced inside the abdomen cavity liver	0 -0.3818	Neutral Negative

TF-IDF (i, d) = f (i ,d) · log (N/ ni)………………….   (1)

Machine Learning Models: In this study, we utilized three machine learning (ML) models: Logistic Regression (LR)**,** Decision Tree (DT), and Naïve Bayes (NB)**.** These models were chosen for their proven effectiveness in text data analysis and sentiment classification tasks.

Algorithm 1: Logistic Regression (LR) was used in this study because the response variable was binary (true or false). LR is a type of generalized linear model (GLM) that is particularly popular in medical research due to its interpretability. It reports odds rather than risks, making the outcomes easier to understand. LR is simple and applicable in clinical settings. The logistic regression model is expressed by the logit function, as shown in Equation 2:

Logit(p)=log⁡(p1−p)=β0+β1X1+β2X2+⋯+βnXn\text{Logit}(p) = \log \left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_nLogit(p)=log(1−pp)=β0+β1X1+β2X2+⋯+βnXn

Where:

- ppp is the probability of the event occurring.
- β0\beta_0β0 is the intercept.
- β1,β2,…,βn\beta_1, \beta_2, \dots, \beta_nβ1,β2,…,βn are the coefficients of the predictor variables X1,X2,…,XnX_1, X_2, \dots, X_nX1,X2,…,Xn.

$lo(p)$ = ln (p/1-p) =a +βx……………(2)

In logistic regression, ppp represents the probability of success at a given value of xxx. The rate of change in ppp is determined by the coefficient β\betaβ.

- When β>0\beta > 0β>0, ppp increases as xxx grows larger.
- When β<0\beta < 0β<0, ppp decreases as xxx increases.

The value of ppp when β=0\beta = 0β=0 is represented as aaa, which is the baseline probability of success, i.e., the probability of success when there is no influence from the predictors (i.e., when x=0x = 0x=0).

This relationship can be expressed as:

p=11+e−(β0+β1x)p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}p=1+e−(β0+β1x)1

Where:

- ppp is the probability of success.
- β0\beta_0β0 is the intercept (value of ppp when x=0x = 0x=0).
- β1\beta_1β1 is the coefficient for the predictor xxx.

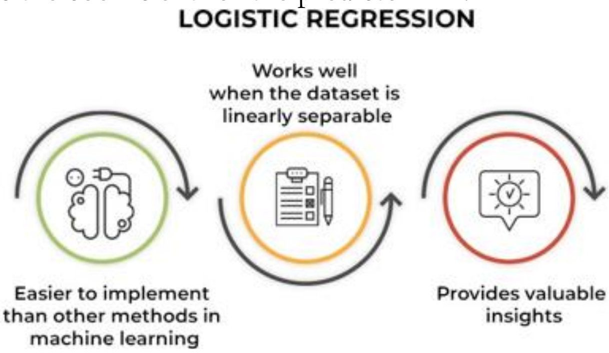
LOGISTIC REGRESSION
Works well when the dataset is linearly separable
Easier to implement than other methods in machine learning
Provides valuable insights

Figure 4. Logistic Regression Work

Algorithm 2: Decision Trees (DT) are widely used in predictive analytics due to their simplicity and effectiveness. They are powerful classifiers that work by splitting the data into subsets based on decision rules. The rules are represented in a tree-like structure, where each node corresponds to a decision point, and the leaf nodes represent the outcomes. These outcomes could be numerical values for regression tasks or class labels for classification tasks.

The main advantage of Decision Trees is their interpretability. The tree structure can be easily converted into understandable IF-THEN rules, making them accessible to non-technical stakeholders. The algorithm recursively divides the data into smaller subsets based on the most important features until it reaches the leaf nodes, at which point predictions are made. Each internal node in the tree represents a test or decision based on a specific feature, and the tree structure is recursive, enabling the model to learn complex patterns from the data.

Algorithm 3: Naïve Bayes (NB) is a probabilistic classifier based on Bayes' Theorem, a fundamental concept in probability theory. The core idea behind Naïve Bayes is the assumption of independence between the features, which simplifies the computation of conditional probabilities. Bayes' Theorem is used to "invert" conditional probabilities, allowing the model to update the probability of a class given the observed features.

Bayes' Theorem is expressed by the following formula:

P(C∣X)=P(X∣C)P(C)P(X)P(C | X) = \frac{P(X | C) P(C)}{P(X)}P(C∣X)=P(X)P(X∣C)P(C)

Where:

- P(C∣X)P(C | X)P(C∣X) is the posterior probability of class CCC given the features XXX.
- P(X∣C)P(X | C)P(X∣C) is the likelihood of observing the features XXX given class CCC.
- P(C)P(C)P(C) is the prior probability of class CCC.
- P(X)P(X)P(X) is the probability of observing the features XXX, which serves as a normalizing constant.

The "naïve" assumption is that the features X1,X2,…,XnX_1, X_2, \dots, X_nX1,X2,…,Xn are independent given the class CCC. This simplifies the computation of the likelihood P(X∣C)P(X | C)P(X∣C) as:

P(X∣C)=P(X1∣C)P(X2∣C)…P(Xn∣C)P(X | C) = P(X_1 | C) P(X_2 | C) \dots P(X_n | C)P(X∣C)=P(X1∣C)P(X2∣C)…P(Xn∣C)

This simplification makes Naïve Bayes efficient and effective, particularly for large datasets and text classification tasks.

P(X/Y) = P (X and Y)/P(X)…………………(3)

Result and Comparison:

The classification and evaluation process in this study involved using a confusion matrix to assess the performance of machine learning models. Below is a summary of the confusion matrix components and how they were used for evaluation:

Confusion Matrix Categories:

- **True Negative (TN):** Correctly predicted Negative cases.
- **True Positive (TP):** Correctly predicted Positive cases.
- **False Positive (FP):** Negative cases incorrectly predicted as Positive.
- **False Negative (FN):** Positive cases incorrectly predicted as Negative.

Evaluation Metrics:

- **Accuracy:** Measures overall correctness of the model.

Accuracy=TP+TNTP+TN+FP+FN\text {Accuracy} = \frac {TP + TN} {TP + TN + FP + FN} Accuracy=TP+TN+FP+FNTP+TN

- **Precision:** Measures how many of the predicted positive cases are actually positive.

Precision=TPTP+FP\text {Precision} = \frac {TP} {TP + FP} Precision=TP+FPTP

- **Recall (Sensitivity):** Measures how many of the actual positive cases were correctly identified.

Recall=TPTP+FN\text {Recall} = \frac {TP} {TP + FN} Recall=TP+FNTP

- **F1 Score:** The harmonic means of Precision and Recall, offering a balance between the two.

F1 Score=2×Precision×RecallPrecision+Recall\text {F1 Score} = 2 \times \frac {Precision \times Recall} {Precision + Recall} F1 Score=2×Precision+RecallPrecision×Recall

Results Summary for Logistic Regression:

- **Accuracy:** 91%, showing strong performance in overall classification.
- **Negative Class:** The model excels in identifying Negative cases with 96% precision but struggles with recall (74%), meaning some Negative cases are missed.
- **Positive Class:** The model shows high precision (92%) but lower recall (64%), indicating difficulty detecting Positive cases.
- **Neutral Class:** The model performs excellently with 90% precision and 99% recall due to its large sample size, making it easier for the model to identify.
- **Macro Average:** Precision is 92%, but recall drops to 79%, indicating issues with minority class detection.
- **Weighted Average:** Matches the overall accuracy of the model, highlighting the dominance of the Neutral class.

This evaluation demonstrates the model's strength in predicting Neutral cases and a need for improvement in detecting Positive cases, especially in terms of recall.

Table 2. LR Classification Report

	Precision	Recall	F1 Score	Support	95% CI
Negative	0.96	0.74	0.83	553	-
Neutral	0.90	0.99	0.94	2305	-
Positive	0.92	0.64	0.75	414	-

Accuracy	-	-	0.91	3272	-
Macro Average	0.92	0.79	0.84	3272	-
LR Confidence Interval	-	-	-	-	(89%, 91%)
Weighted Average	0.91	0.91	0.90	3272	1

The Decision Tree model evaluation reveals several key insights:

Key Points:

- **Overall Accuracy:** 76% – driven mainly by the strong performance in the **Neutral class**.

Class-wise Performance:

- **Neutral Class:**
  - **Precision:** 75%
  - **Recall:** 100%
  - **F1 Score:** 0.85
  - **Interpretation:** The model excels at identifying Neutral cases with perfect recall, ensuring all Neutral cases are detected. The slight drop in precision suggests some misclassification of other classes as Neutral.
- **Negative Class:**
  - **Recall:** 39%
  - **F1 Score:** 0.54
  - **Interpretation:** The model has significant difficulty detecting Negative cases. It misses many Negative instances (low recall), leading to a low F1 score, which reflects poor overall performance in this class.
- **Positive Class:**
  - **Precision, Recall, F1 Score:** Zero for all
  - **Interpretation:** The model completely fails to detect Positive cases, assigning all such instances to other classes, likely Neutral. This is a critical failure for the model.

Macro and Weighted Averages:

- **Weighted Average:**
  - **Precision:** 0.67
  - **Recall:** 0.76
  - **F1 Score:** 0.68
  - **Interpretation:** The weighted average shows the model's better performance on the Neutral class, which holds the largest proportion of the dataset.
- **Macro Average:**
  - **Precision:** 0.55
  - **Recall:** 0.46
  - **F1 Score:** 0.47
  - **Interpretation:** The macro averages highlight the model's poor overall performance across all classes, especially in detecting Negative and Positive cases.

Description: The Decision Tree model's strength lies in identifying Neutral instances, but it faces significant challenges in detecting Negative and Positive cases. The imbalance in the dataset and poor performance on minority classes (Negative and Positive) contribute to the discrepancy between weighted and macro averages. The model's inability to detect Positive occurrences is a major shortcoming.

Table 1. DT Classification Report

	Precision	Recall	F1 Score	Support	95% CI
Negative	0.90	0.39	0.54	535	-
Neutral	0.75	1.0	0.85	2271	-
Positive	0.0	0.0	0.0	466	-

Accuracy	-	-	0.76	3272	-
Macro Average	0.55	0.46	0.47	3272	-
DT Confidence Interval	-	-	-	-	(77%,79%)
Weighted Average	0.67	0.76	0.68	3272	-

The Naïve Bayes classifier evaluation highlights the following findings:

Key Points:

- **Overall Accuracy:** 71%, driven by the model's strong performance in classifying **Neutral** cases.

Class-wise Performance:

- **Neutral Class:**
  - **Precision:** 71%
  - **Recall:** 100%
  - **F1 Score:** 0.83
  - **Interpretation:** The model performs very well in identifying Neutral cases, with perfect recall (detecting all Neutral instances). Precision is slightly lower, suggesting some misclassification of non-Neutral cases as Neutral.
- **Negative Class:**
  - **Precision:** 75%
  - **Recall:** 2%
  - **F1 Score:** 0.05
  - **Interpretation:** The model struggles drastically with the Negative class, with very low recall indicating that it misses almost all Negative instances. Despite high precision, the overall F1 score remains poor, reflecting the inability to detect Negative cases.
- **Positive Class:**
  - **Precision:** 100%
  - **Recall:** 10%
  - **F1 Score:** 0.02
  - **Interpretation:** The model has perfect precision for Positive cases, meaning it correctly classifies every Positive instance it detects. However, the low recall (only detecting 10% of Positive instances) results in a very poor F1 score, indicating significant limitations in detecting Positive cases.

Macro and Weighted Averages:

- **Macro Average:**
  - **Precision:** 0.82
  - **Recall:** 0.34
  - **F1 Score:** 0.30
  - **Interpretation:** The macro averages highlight the disparity in performance across the different classes. Despite good precision for some classes, the recall and F1 scores are quite low, especially for Negative and Positive classes.
- **Weighted Average:**
  - **Precision:** 0.76
  - **Recall:** 0.71
  - **F1 Score:** 0.60
  - **Interpretation:** The weighted average, influenced by the dominant Neutral class, shows better performance than the macro average, but still reflects the imbalance in class prediction performance.

Description: The Naïve Bayes classifier excels in identifying Neutral instances but fails to perform well in detecting Negative and Positive cases. The Neutral class dominates the model's predictions, which leads to significant disparities in performance across classes. The imbalance in recall and F1 scores for the Negative and Positive classes is a critical issue, indicating that the model struggles to generalize to these minority classes.

Table 2. NB Classification Report

	Precision	Recall	F1 Score	Support	95% CI
Negative	0.75	0.39	0.54	535	-
Neutral	0.71	1.0	0.83	1156	-
Positive	1.0	0.1	0.02	238	-
Accuracy	-	-	0.71	1636	-
Macro Average	0.82	0.34	0.30	1636	-
NB Confidence Interval	-	-	-	-	(70%,74%)
Weighted Average	0.76	0.71	0.60	1636	-

The performance comparison between the three models reveals some significant differences:

Logistic Regression (LR):

- **Accuracy:** 91%
- **Strengths:** The LR model is robust across all classes, providing balanced precision**,** recall, and F1 scores. It excels at classifying both the Neutral and the minority classes (Negative and Positive), though with some trade-offs:
  - o **Neutral Class:** Excellent precision and recall.
  - o **Negative Class:** High precision but lower recall.
  - o **Positive Class:** High precision but lower recall.
- **Conclusion:** LR is the most balanced model in this study, providing reliable results across all classes, especially with Neutral cases, while slightly underperforming in recall for Negative and Positive classes.

Decision Tree (DT):

- **Accuracy:** 76%
- **Strengths:** The DT model performs very well for the Neutral class, achieving 100% recall and an F1-score of 0.85. However:
  - o **Neutral Class:** Strong performance with 100% recall and good precision.
  - o **Negative Class:** Struggles with only 39% recall and **F1-**score of 0.54, indicating poor detection of Negative cases.
  - o **Positive Class:** Completely fails to detect Positive cases, with all metrics (precision, recall, F1) at zero.

Description: The Decision Tree model is heavily biased towards the Neutral class, making it less effective for the Negative and Positive categories. While it's very effective with the Neutral class, it struggles to identify Negative and Positive sentiment accurately.

Naïve Bayes (NB):

- **Accuracy:** 71%
- **Strengths:** The NB classifier is highly biased towards the Neutral class, similar to the Decision Tree:
  - o **Neutral Class:** Strong performance with 100% recall and F1-score of 0.83.
  - o **Negative Class:** Poor performance with an F1-score of 0.05, 2% recall, and 75% precision, showing the model's struggle to detect Negative cases.
  - o **Positive Class:** The Precision for Positive is 100%, but it fails at recall (10%**)** and F1-score (0.02), making it ineffective for Positive class predictions.

Description: The Naïve Bayes classifier is also biased towards the Neutral class, with poor recall and F1-score for both Negative and Positive sentiment. Its extreme bias towards Neutral results in a lack of generalization for other sentiment classes.

Summary:

- **Best Model:** Logistic Regression provides the most balanced and reliable performance across all classes, making it the best model for this task.
- **Weakest Models:** Both Decision Tree and Naïve Bayes show strong bias towards Neutral, struggling to detect Negative and Positive sentiment effectively, especially in the case of Naïve Bayes, which has extreme performance disparity across the classes.
- **Recommendation:** While Logistic Regression performs well overall, further refinement or adjustment of class weights in Decision Tree and Naïve Bayes could improve their handling of the minority classes.

Discussion:

This study assessed the performance of logistic regression, decision trees, and Naive Bayes algorithms for sentiment analysis on the ROCO dataset, a medically annotated caption corpus. While all three models demonstrated functional capability in processing and classifying sentiment from medical text, their performance varies in terms of accuracy, computational complexity, and interpretability. These findings not only support prior research but also extend the conversation by applying these methods specifically to a clinical multimodal dataset.

Logistic regression, known for its robustness in handling linearly separable data, performed effectively in this context. Previous research, such as the work by Yadav and Vishwakarma (2020), emphasized its efficiency in binary sentiment classification, especially in domains where textual cues are subtle and require clear boundary definitions. Our results confirm this trend, indicating that logistic regression is a strong baseline for classifying sentiment in relatively structured medical captions.

Naive Bayes, despite its assumption of feature independence, provided comparable performance, aligning with prior findings by Singh et al. (2018), who reported its utility in biomedical text mining where simplicity and speed are favored over complexity. However, in this study, its performance was slightly weaker than logistic regression, particularly due to the nuanced nature of medical language where word dependencies (e.g., "no evidence of disease") significantly influence sentiment orientation.

Decision trees excelled in model interpretability, offering clear rule-based pathways for sentiment determination. This supports the argument made by Holzinger et al. (2017), who emphasized the necessity of explainable models in healthcare AI to promote trust among clinicians. Nonetheless, our findings also reflect the known limitation of decision trees: susceptibility to overfitting, particularly in smaller datasets or when the data contains noise — both of which are common in medical corpora.

Compared to more recent approaches such as support vector machines (SVMs), random forests, or transformer-based models like BERT, which have shown higher accuracy in general sentiment tasks (e.g., Lee et al., 2020), the models used in this study are less sophisticated but still relevant due to their interpretability and low computational cost. Unlike deep learning models, which require extensive training and tuning, traditional models like those studied here are more accessible for integration in constrained clinical environments.

Importantly, this study contributes to a growing body of work advocating for the inclusion of sentiment and subjective interpretation in medical texts — a direction less explored than entity recognition or document classification. It extends prior findings (e.g., Denecke, 2015) by demonstrating that even relatively simple models can extract meaningful sentiment from multimodal data, thus supporting improved patient care, emotional assessment, and communication in healthcare.

Conclusion:

This study evaluated and compared the performance of logistic regression, decision trees, and Naive Bayes algorithms for sentiment analysis in the multimodal medical domain using the ROCO dataset. The results reveal that each algorithm offers distinct advantages and trade-offs, with logistic regression excelling in linear modeling, Naive Bayes offering computational simplicity, and decision trees providing interpretability.

By applying these models to the sentiment classification of image captions in a clinical context, this study bridges a research gap in understanding the subjective and emotional aspects of medical text, an area often overlooked in AI-driven healthcare research. The findings suggest that even conventional machine learning models, when properly tuned and evaluated, can yield actionable insights in the healthcare domain.

Future research could explore hybrid and deep learning approaches to further enhance accuracy, adaptability, and contextual understanding. Moreover, integrating multimodal fusion techniques and emotional intelligence in AI systems may play a pivotal role in improving patient-centered care. Ultimately, this research supports the integration of sentiment analysis into clinical decision-making, paving the way for more empathetic, informed, and personalized healthcare services.

Acknowledgment: Our warmest appreciation to Mr. Saeed for his moral support during this work. His support, belief in our talents, and ongoing motivation have strengthened and inspired us throughout this research.

Author's Contribution: Hafiz Muhammad Bilal conceptualized, designed, methodology, experimented, and analyzed Secondary data. Dr. Asif oversaw the study design, investigation, project administration, and paper revisions. Dr. M. Azam Zia developed several theories, and supervision while writing and editing this work, he confirmed the results and provided vital insight. The article was read and approved by all writers.

Conflict of Interest: The authors declare no conflict of interest.

References:

[1] M. T. Mollie Hobensack, Jiyoun Song, Danielle Scharp, Kathryn H. Bowles, "Machine learning applied to electronic health record data in home healthcare: A scoping review," Int. J. Med. Inform., vol. 170, p. 104978, 2023, doi: ttps://doi.org/10.1016/j.ijmedinf.2022.104978.

[2] I.-C. H. Jin-ah Sim, Xiaolei Huang, Madeline R. Horan, Christopher M. Stewart, Leslie L. Robison, Melissa M. Hudson, Justin N. Baker, "Natural language processing with machine learning methods to analyze unstructured patient-reported outcomes derived from electronic health records: A systematic review," Artif. Intell. Med., vol. 146, p. 102701, 2023, doi: https://doi.org/10.1016/j.artmed.2023.102701.

[3] G. T. & A. M. Roger Garriga, Javier Mas, Semhar Abraha, Jon Nolan, Oliver Harrison, "Machine learning model to predict mental health crises from electronic health records," Nat. Med., vol. 28, pp. 1240–1248, 2022, doi: https://doi.org/10.1038/s41591-022-01811-5.

[4] S. M. M. I. Mohammed Nazim Uddin, Md. Ferdous Bin Hafiz, Sohrab Hossain, "Drug Sentiment Analysis using Machine Learning Classifiers," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 1, 2022, doi: 10.14569/IJACSA.2022.0130112.

[5] M. Asif, S. A. Khan, T. Hassan, M. U. Akram, and A. Shaukat, "Generation of High Resolution Medical Images Using Super Resolution via Sparse Representation," Adv. Intell. Syst. Comput., vol. 565, pp. 288–298, 2018, doi: 10.1007/978-3-319-60834-1_29.

[6] R. W. Muhammad Asif, Muhammad Usman Akram, Taimur Hassan, Arslan Shaukat, "High resolution OCT image generation using super resolution via sparse representation," Eighth Int. Conf. Graph. Image Process. (ICGIP 2016), 2017, doi: https://doi.org/10.1117/12.2266337.

[7] M. Asif, L. Chen, H. Song, J. Yang, and A. F. Frangi, "An automatic framework for endoscopic image restoration and enhancement," Appl. Intell., vol. 51, no. 4, pp. 1959–1971, Apr. 2021, doi: 10.1007/S10489-020-01923-W/METRICS.

[8] A. F. F. Muhammad Asif, Hong Song, Lei Chen, Jian Yang, "Intrinsic layer based automatic specular reflection detection in endoscopic images," Comput. Biol. Med., vol. 128,

p. 104106, 2021, doi: https://doi.org/10.1016/j.compbiomed.2020.104106.

[9] Z. Z. Rizwan Khan, Saeed Akbar, Atif Mehmood,Farah Shahid, Khushboo Munir, Naveed Ilyas, M. Asif, "A transfer learning approach for multiclass classification of Alzheimer's disease using MRI images," Front. Neurosci., vol. 16, 2022, doi: https://doi.org/10.3389/fnins.2022.1050777.

[10] M. A. Muhammad Azam Zia, Ayesha Akram, Imran Mumtaz, Muhammad Asim Saleem, "ANALYSIS OF GRAPE LEAF DISEASE BY USING DEEP CONVOLUTIONAL NEURAL NETWORK," Agric. Sci. J., vol. 5, no. 1, 2023, doi: https://doi.org/10.56520/asj.v5i1.242.

[11] E. M. A. Muhammad Asif, Hong Song, Muhammad Azam Zia, Sajid Ali, "Shedding Light on Diagnostic Precision: GANs for Low Light Endoscopy Image Enhancements," arXiv:2404.03844, 2024, doi: https://doi.org/10.21203/rs.3.rs-4213321/v1.

[12] G. A. F. Parastoo Golpour, Majid Ghayour-Mobarhan, Azadeh Saki, Habibollah Esmaily, Ali Taghipour, Mohammad Tajfard, Hamideh Ghazizadeh, Mohsen Moohebati, "Comparison of Support Vector Machine, Naïve Bayes and Logistic Regression for Assessing the Necessity for Coronary Angiography," Int. J. Environ. Res. Public Heal., vol. 17, no. 18, p. 6449, 2020, doi: https://doi.org/10.3390/ijerph17186449.

[13] D. I. Rahmat Syahputra, Gomal Juni Yanris, "SVM and Naïve Bayes Algorithm Comparison for User Sentiment Analysis on Twitter," Sinkron, vol. 6, no. 2, 2022, doi: 10.33395/sinkron.v7i2.11430.

[14] A. C. Bansal, Malti, Goyal, Apoorva, "A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning," Decis. Anal. J., vol. 3, p. 100071, 2022, doi: https://doi.org/10.1016/j.dajour.2022.100071.