*International Journal of Recent Research and Review, Vol. XIX, Issue 3, September 2026*

ISSN 2277 – 8322

# MACHINE LEARNING-BASED OPINION MINING FOR HEALTHCARE FEEDBACK ANALYSIS

Soniya Kharra

Assistant Professor

Jagannath University, Jaipur, Rajasthan, India

**Abstract-** The growing use of digital healthcare platforms has increased the availability of patient feedback, creating opportunities to assess satisfaction and identify areas for service improvement. This study presents a machine learning-based opinion mining framework for classifying healthcare feedback into positive, negative, and neutral sentiments. The methodology involves text preprocessing, including data cleaning, tokenization, stop-word handling, and text normalization, followed by feature extraction using Term Frequency–Inverse Document Frequency (TF-IDF). Four supervised machine learning algorithms—Logistic Regression, Linear Support Vector Machine (SVM), Multinomial Naïve Bayes, and Random Forest—are comparatively evaluated using accuracy, precision, recall, macro F1-score, and confusion matrix analysis. The tentative results indicate that Linear SVM achieves the highest performance, with 91.6% accuracy and a 91.0% macro F1-score, compared with the other evaluated classifiers. The framework also considers feedback themes such as staff behaviour, waiting time, cleanliness, communication, and facilities to identify potential areas of patient satisfaction and dissatisfaction. The proposed approach demonstrates the potential of TF-IDF-based machine learning techniques to support systematic healthcare feedback analysis and data-driven service improvement. However, the reported performance metrics and thematic findings are illustrative and must be validated using actual experimental data before being reported as final research findings.

**Keywords**: Opinion Mining, Machine Learning, Healthcare Feedback Analysis, Sentiment Analysis, Patient Satisfaction, TF-IDF, Support Vector Machine, Text Classification.

## I. INTRODUCTION

### 1.1 Background of the Study

The rapid advancement of information technology, artificial intelligence (AI), and data

analytics has significantly transformed the healthcare sector. Healthcare organizations increasingly rely on digital platforms to manage patient information, deliver medical services, and collect feedback regarding patient experiences. Patients share their opinions through hospital websites, online review portals, healthcare applications, social media platforms, and digital surveys. These feedback sources contain valuable information about the quality of medical services, healthcare professionals' behaviour, waiting times, hospital facilities, treatment experiences, and overall patient satisfaction. Consequently, analyzing patient feedback has become an important component of healthcare quality assessment and continuous service improvement. Traditional approaches to feedback analysis generally involve manual reading, categorization, and interpretation of patient comments. Although these approaches may be effective for small datasets, they become inefficient when healthcare organizations receive large volumes of textual feedback. Manual assessment is also susceptible to subjective interpretation, inconsistency, and delays in identifying emerging patient concerns. Therefore, automated techniques are required to process unstructured healthcare feedback efficiently and extract meaningful information that can support administrative and clinical service improvements. Opinion mining, commonly known as sentiment analysis, is a natural language processing (NLP) technique used to identify and interpret opinions, emotions, and attitudes expressed in textual data. In healthcare, opinion mining can be applied to classify patient feedback into positive, negative, and neutral sentiments. Machine learning algorithms enable these classifications by learning patterns from previously labelled textual data and applying the learned patterns to unseen feedback. The integration of machine learning and opinion mining therefore provides an effective approach to understanding patient experiences and identifying opportunities for improving healthcare services.

### 1.2 Concept of Opinion Mining in Healthcare

Opinion mining refers to the computational process of identifying opinions and determining the sentiment expressed in written or spoken language. It involves analyzing textual content to determine whether a statement communicates a favourable, unfavourable, or neutral opinion about a particular subject. In healthcare applications, the subject may include hospital services, medical staff, nursing care, appointment scheduling, diagnostic facilities, treatment procedures, or patient

communication. For example, a patient comment such as “The doctor explained the treatment clearly, and the staff were very supportive” generally expresses positive sentiment. In contrast, a statement such as “The waiting time was excessive, and the staff did not respond to my concerns” indicates negative sentiment. A statement that merely describes an appointment date or hospital location may be classified as neutral, depending on the classification criteria and contextual information. Healthcare feedback is particularly challenging to analyze because patient opinions frequently contain medical terminology, abbreviations, spelling errors, emotional expressions, and context-dependent language. The same word may communicate different meanings depending on the surrounding sentence. Sarcasm, negation, mixed sentiments, and multilingual expressions can further complicate automated classification. Consequently, healthcare opinion mining requires appropriate text preprocessing, feature representation, and classification techniques to improve the reliability of sentiment identification.

### 1.3 Role of Machine Learning in Sentiment Analysis

Machine learning is a branch of artificial intelligence that enables computer systems to identify patterns in data and make predictions without relying exclusively on manually defined rules. In sentiment analysis, supervised machine learning algorithms are trained using labelled feedback containing predefined sentiment categories. During training, the algorithms learn associations between textual features and sentiment labels. The trained models are subsequently evaluated on unseen data to determine their ability to classify new patient comments. Several machine learning algorithms are suitable for healthcare feedback classification. Logistic Regression is commonly used for binary and multiclass classification and provides a useful baseline for text classification tasks. Support Vector Machine (SVM) can effectively handle high-dimensional textual feature spaces and is frequently considered for sentiment classification. Multinomial Naïve Bayes offers a computationally efficient probabilistic approach, particularly for text represented using word-frequency features. Random Forest combines multiple decision trees and can capture certain nonlinear relationships in structured feature representations. Textual information must be converted into numerical features before it can be processed by conventional machine learning algorithms. Term Frequency–Inverse Document Frequency (TF-IDF) is a widely used technique that

assigns weights to words based on their frequency within a document and their distribution across the document collection. This representation helps emphasize terms that are informative for distinguishing between different feedback categories. By combining TF-IDF with suitable classification algorithms, a machine learning-based system can automatically analyze patient feedback and produce sentiment predictions.

### 1.4 Importance of Healthcare Feedback Analysis

Patient feedback provides healthcare organizations with an additional source of information about service quality and patient experience. Systematic analysis of this information can help identify recurring problems, evaluate patient satisfaction, and support improvements in hospital operations. Positive feedback may reveal service characteristics that patients value, whereas negative feedback may indicate areas requiring administrative attention. For example, repeated complaints about appointment delays may indicate the need to review scheduling procedures. Negative comments concerning staff communication may highlight opportunities for improving patient interaction and information delivery. Feedback about cleanliness, waiting facilities, billing procedures, and diagnostic services can also assist administrators in identifying operational weaknesses. Automated sentiment analysis can reduce the effort required to review large collections of patient comments. It can also support the periodic monitoring of feedback trends and comparisons across departments, facilities, or time periods, provided that the data are sufficiently representative and consistently collected. However, sentiment classification should be treated as a decision-support mechanism rather than a direct measure of clinical quality. Patient opinions reflect personal experiences and may not independently establish the effectiveness, safety, or clinical appropriateness of medical treatment.

## II. RESEARCH METHODOLOGY

### 2.1 Research Design

The present study adopts an experimental and comparative research methodology to develop and evaluate a machine learning-based opinion mining framework for healthcare feedback analysis. The proposed framework focuses on automatically classifying healthcare-related textual feedback into three sentiment categories: positive, negative, and neutral. The methodology integrates data collection, text preprocessing, feature extraction, machine

learning model development, sentiment classification, and performance evaluation. A comparative approach is employed to examine the effectiveness of different machine learning algorithms under consistent experimental conditions. The overall objective is to identify a suitable classification model for analyzing patient opinions and extracting meaningful information from healthcare feedback.

### 2.2 Dataset Collection and Description

The first stage involves collecting healthcare-related textual feedback from a publicly available dataset, an authorized hospital feedback repository, or another ethically approved source. The dataset may contain patient reviews concerning hospital services, medical staff behaviour, waiting time, treatment experiences, nursing care, cleanliness, appointment management, and overall satisfaction. Each record should ideally include a feedback text and an associated sentiment label. The sentiment labels may be obtained from existing annotations or assigned through a predefined annotation procedure.

For a supervised machine learning approach, the dataset must contain sufficient labelled examples representing the sentiment categories. The quality and representativeness of the dataset are important because model performance depends on the volume, diversity, and accuracy of the available data. Before analysis, the dataset is examined for duplicate records, missing values, irrelevant entries, and inconsistent labels. Patient identifiers and other sensitive information must be removed or appropriately protected before processing. The dataset source, number of records, class distribution, language, and inclusion criteria should be documented to ensure reproducibility.

### 2.3 Data Preprocessing

Healthcare feedback is generally available in an unstructured textual format and may contain spelling errors, punctuation, HTML elements, abbreviations, and irrelevant symbols. Therefore, preprocessing is performed to improve text consistency and prepare the data for feature extraction. The preprocessing stage consists of the following operations:

1. **Data cleaning:** Removal of duplicate records, empty feedback entries, unnecessary HTML tags, and irrelevant symbols.
2. **Text normalization:** Conversion of text into a consistent format, such as lowercase, where appropriate.

3. **Tokenization:** Division of sentences into individual words or tokens for further processing.

4. **Stop-word handling:** Removal of non-informative words when appropriate. Negation terms such as “not” and “never” must be retained because they can change the meaning of a sentiment-bearing statement.

5. **Linguistic normalization:** Application of stemming or lemmatization, where beneficial, to reduce variations of words to common forms.

6. **Label verification:** Examination of sentiment labels to minimize inconsistencies and improve the reliability of supervised learning.

The same preprocessing rules must be applied consistently to training and testing data. Care is taken to avoid removing meaningful healthcare terminology or emotional expressions that contribute to sentiment classification.

### 2.4 Feature Extraction Using TF-IDF

After preprocessing, textual feedback is transformed into numerical representations using Term Frequency–Inverse Document Frequency (TF-IDF). This technique estimates the importance of a term within a particular feedback document relative to its occurrence throughout the dataset. TF-IDF is suitable for conventional machine learning algorithms because it represents textual documents as numerical feature vectors.

The TF-IDF weight of a term can be expressed as:

[
\operatorname{TFIDF}(t,d)=\operatorname{TF}(t,d)\times\operatorname{IDF}(t)
]

where (\operatorname{TF}(t,d)) represents the frequency of term (t) in document (d), and (\operatorname{IDF}(t)) represents the inverse document frequency of the term across the document collection.

The resulting feature matrix provides a numerical representation of healthcare feedback for model training. Unigrams and bigrams may be considered to capture individual words and short phrases, such as “friendly staff,” “long waiting,” and “not satisfied.” Vocabulary selection and other TF-IDF parameters are determined using training data only to prevent information leakage from the test dataset.

### 2.5 Machine Learning Model Development

Four supervised machine learning algorithms are selected for comparative evaluation: Logistic Regression, Support Vector Machine, Multinomial Naïve Bayes, and Random Forest. These algorithms provide different approaches to learning relationships between textual features and sentiment labels.

**Logistic Regression:** This algorithm estimates the probability that a feedback record belongs to a particular sentiment category. It serves as a computationally efficient baseline for text classification and can provide interpretable feature-weight information.

**Support Vector Machine (SVM):** SVM identifies decision boundaries that separate sentiment categories within the feature space. A linear SVM is considered suitable for high-dimensional sparse text representations and can provide strong classification performance.

**Multinomial Naïve Bayes:** This probabilistic classifier estimates sentiment categories using the distribution of textual features. It is computationally efficient and commonly used as a baseline for text classification tasks.

**Random Forest:** This ensemble learning method combines predictions from multiple decision trees. It is included to examine whether an ensemble-based approach can effectively classify feedback using the selected feature representation. Its performance may depend on dataset characteristics and parameter settings.

Each model is trained using the same training dataset and evaluated using the same test dataset. Appropriate hyperparameter tuning and stratified cross-validation may be performed within the training data to select suitable model configurations. The final test set is reserved for independent evaluation.

### 2.6 Sentiment Classification Process

The sentiment classification stage uses the trained models to predict the sentiment category of each feedback record. The output consists of positive, negative, or neutral labels. Positive feedback generally indicates satisfaction or favourable experiences, negative feedback indicates dissatisfaction or unfavourable experiences, and neutral feedback contains information without a clearly identifiable positive or negative orientation according to the annotation guidelines.

The complete classification process begins with the input of healthcare feedback, followed by preprocessing and TF-IDF transformation. The resulting feature vectors are supplied to the trained classification model, which generates the predicted sentiment label. These predictions

can then be aggregated to determine the overall sentiment distribution within the dataset.

Where sufficient information is available, feedback may also be grouped into service-related themes, such as staff behaviour, waiting time, cleanliness, communication, and facilities. Such thematic analysis should be performed using a clearly defined method and validated where necessary. Sentiment classification alone does not automatically establish the exact cause of patient satisfaction or dissatisfaction.

### 2.7 Performance Evaluation Metrics

The performance of the proposed models is assessed using standard classification metrics derived from the confusion matrix. The evaluation considers correct and incorrect predictions across the sentiment categories.

**Accuracy** measures the proportion of correctly classified feedback records among all evaluated records.

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

**Precision** measures the proportion of predicted positive cases that are correctly classified.

$$\text{Precision}=\frac{TP}{TP+FP}$$

**Recall** measures the proportion of actual positive cases correctly identified.

$$\text{Recall}=\frac{TP}{TP+FN}$$

**F1-score** represents the harmonic mean of precision and recall.

$$F1=2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$$

Here, (TP), (TN), (FP), and (FN) denote true positives, true negatives, false positives, and false negatives, respectively. For multiclass sentiment classification, these metrics are calculated using a one-versus-rest approach for each category. Macro-averaged precision, recall, and F1-score are reported to provide equal weight to each class, particularly when the dataset is imbalanced. A confusion matrix is also used to identify the categories most frequently misclassified by each model.

### 2.8 Comparative Analysis and Interpretation

The results obtained from the four machine learning algorithms are compared using the same evaluation dataset and performance metrics. The comparison examines overall accuracy, class-specific precision and recall, macro-F1 score, and the distribution of classification errors. The most suitable model is selected based on the combined evaluation results rather than accuracy alone. Computational efficiency and the ability to identify minority sentiment categories may also be considered.

The sentiment distribution and recurring feedback themes are summarized using tables and graphical representations where appropriate. The analysis aims to identify useful patterns in patient opinions and examine the potential of automated feedback classification to support healthcare service improvement. Any conclusions are restricted to the characteristics and limitations of the dataset used in the study.

## III. RESULTS AND DISCUSSION

The material provided describes the proposed methodology but does not include an actual dataset, model outputs, or metric values. Therefore, valid research results cannot be calculated or claimed without running the four models on labelled healthcare-feedback data. The following Word-ready section is a results template: replace all bracketed values with your experimentally obtained values before submission.

### 3.1 Dataset Summary

A total of **[N]** healthcare feedback records were included after removing duplicate, empty, and irrelevant entries. The dataset was divided into training and testing sets using a stratified split of **[80%]** and **[20%]**, respectively, so that positive, negative, and neutral feedback remained represented in both subsets.

| Sentiment class | Number of records | Percentage |
|---|---|---|
| **Positive** | **550** | **55%** |
| **Negative** | **300** | **30%** |
| **Neutral** | **150** | **15%** |
| **Total** | **1,000** | **100%** |

**Training set: 800 records. Testing set: 200 records.**

**Table 1. Sentiment distribution of healthcare feedback**

### 3.2 Model Performance

The four TF-IDF-based machine-learning classifiers were evaluated using accuracy, precision, recall, and macro F1-score. The best model should be selected using macro F1-score together with accuracy, because macro averaging gives equal importance to positive, negative, and neutral classes.

| Model | Accuracy (%) | Precision (%) | Recall (%) | Macro F1-score (%) |
|---|---|---|---|---|
| **Logistic Regression** | **89.2** | **88.7** | **88.2** | **88.4** |
| **Linear SVM** | **91.6** | **91.1** | **91.0** | **91.0** |
| **Multinomial Naïve Bayes** | **86.4** | **85.9** | **85.1** | **85.5** |
| **Random Forest** | **88.1** | **87.6** | **87.0** | **87.3** |

**Tentative best model: Linear SVM, with 91.6% accuracy and a macro F1-score of 91.0%.**

**Results interpretation template:**

Among the evaluated classifiers, **[best model]** achieved the highest accuracy of **[x]%** and macro F1-score of **[y]%**. This indicates that the model provided the most balanced classification across the three sentiment classes. **[weakest model]** showed comparatively lower performance, possibly because its learning assumptions were less suitable for the sparse, high-dimensional TF-IDF representation of healthcare feedback.

**Table 2. Accuracy comparison of machine-learning models**

### 3.3 Confusion Matrix

Use the confusion matrix for the best-performing model. Each row represents the actual sentiment category, while each column represents the predicted category. Higher values on the main diagonal indicate better classification performance.

**Illustrative confusion matrix for Linear SVM on the assumed 200-record test set:**

| Actual / Predicted | Positive | Negative | Neutral |
|---|---|---|---|
| **Positive** | **106** | **3** | **1** |
| **Negative** | **4** | **54** | **2** |
| **Neutral** | **3** | **4** | **23** |

**Figure 3. Confusion-matrix heatmap**

In Word, insert a 3 × 3 table using the above values. Apply darker shading to higher values and lighter shading to lower values. Title it: *"Confusion Matrix of [Best Model] for Healthcare Feedback Classification."*

### 3.4 Feedback Themes

After sentiment classification, recurring patient concerns can be grouped into themes such as staff behaviour, waiting time, cleanliness, communication, facilities, and treatment experience. Report these findings only after using a defined keyword, manual-coding, or aspect-extraction method.

ISSN 2277 – 8322

| **Feedback theme** | **Positive mentions** | **Negative mentions** | **Main observation** |
|---|---|---|---|
| **Staff behaviour** | **180** | **35** | **Supportive interactions may improve patient satisfaction.** |
| **Waiting time** | **45** | **120** | **Delays may be an important source of dissatisfaction.** |
| **Cleanliness** | **100** | **25** | **Positive comments may indicate satisfaction with facilities.** |
| **Communication** | **110** | **50** | **Clear explanations may improve patient experience.** |
| **Facilities** | **95** | **40** | **Infrastructure and amenities may influence opinions.** |

**These counts are also illustrative. They assume a separate theme-identification process and should not be presented as observed results.**

**Table 4. Positive and negative feedback by service theme**

## IV. CONCLUSION OF RESULTS

The present study focuses on the application of machine learning techniques for opinion mining and healthcare feedback analysis. The proposed framework integrates text preprocessing, TF-IDF feature extraction, and supervised machine learning algorithms to classify patient feedback into positive, negative, and neutral sentiment categories. The comparative evaluation considers Logistic Regression, Linear Support Vector Machine (SVM), Multinomial Naïve Bayes, and Random Forest using accuracy, precision, recall, and macro F1-score as performance indicators.

Based on the tentative results presented, Linear SVM demonstrated the best performance, achieving an illustrative accuracy of 91.6% and a macro F1-score of 91.0%, followed by Logistic Regression, Random Forest, and Multinomial Naïve Bayes. These results suggest that Linear SVM may be a promising classifier for high-dimensional textual representations generated using TF-IDF. However, these performance values are illustrative and require validation through actual experiments before drawing definitive conclusions.

The feedback theme analysis further highlights the potential importance of staff behaviour, waiting time, cleanliness, communication, and healthcare facilities in understanding patient experiences. In particular, the illustrative findings indicate that waiting time may be an important source of negative feedback, whereas staff behaviour and communication may provide opportunities to strengthen patient satisfaction. Automated opinion mining can help healthcare administrators organize large volumes of textual feedback, identify recurring concerns, and prioritize areas for service improvement.

Overall, the proposed framework provides a systematic approach to transforming unstructured healthcare feedback into meaningful sentiment information. Future research may extend this work by incorporating larger and more diverse datasets, aspect-based sentiment analysis, deep learning, transformer-based language models, and multilingual feedback classification. The integration of these techniques may improve contextual understanding and classification performance. Nevertheless, actual effectiveness must be established through rigorous experimental validation, appropriate privacy safeguards, and evaluation across different healthcare settings. The proposed approach has the potential to support data-driven healthcare management and continuous improvement in patient-centred services.

**References**

[1] Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. Knowledge-Based Systems, 226, Article 107134.

[2] Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. IEEE Intelligent Systems, 28(2), 15–21.

[3] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language

understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171–4186).

[4] Feizollah, A., Lin, C.-Y., O'Malley, L., Thompson, W., Listl, S., & Byrne, M. (2025). The use of natural language processing to interpret unstructured patient feedback on health services: Scoping review. Journal of Medical Internet Research, 27, Article e72853.

[5] Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the International AAAI Conference on Web and Social Media, 8(1), 216–225.

[6] Jelin, E., Storset, L. C., Norman, R. M., Iversen, H. H. H., Ellingsen-Dalskau, L. H., Mæhlum, P., Velldal, E., Øvrelid, L., & Bjertnaes, O. (2025). From words to action? A scoping review on automatic sentiment analysis of patient experience comments from online sources and surveys. BMJ Health & Care Informatics, 32(1), Article e101631.

[7] Khanbhai, M., Anyadi, P., Symons, J., Flott, K., Darzi, A., & Mayer, E. (2021). Applying natural language processing and machine learning techniques to patient experience feedback: A systematic review. BMJ Health & Care Informatics, 28(1), Article e100262.

[8] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. Information, 10(4), Article 150.

[9] Liu, B. (2012). Sentiment analysis and opinion mining. Morgan & Claypool Publishers.

[10] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093–1113.

[11] Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning–based text classification: A comprehensive review. ACM Computing Surveys, 54(3), Article 62.

[12] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning

in Python. Journal of Machine Learning Research, 12, 2825–2830.

[13] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational Linguistics, 37(2), 267–307.

[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30).

[15] Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55, 5731–5780.

[16] Yadollahi, A., Shahraki, A. G., & Zaiane, O. R. (2017). Current state of text sentiment analysis from opinion to emotion mining. ACM Computing Surveys, 50(2), Article 25.

[17] Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. WIREs Data Mining and Knowledge Discovery, 8(4), Article e1253.

[18] Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion, 37, 98–125.