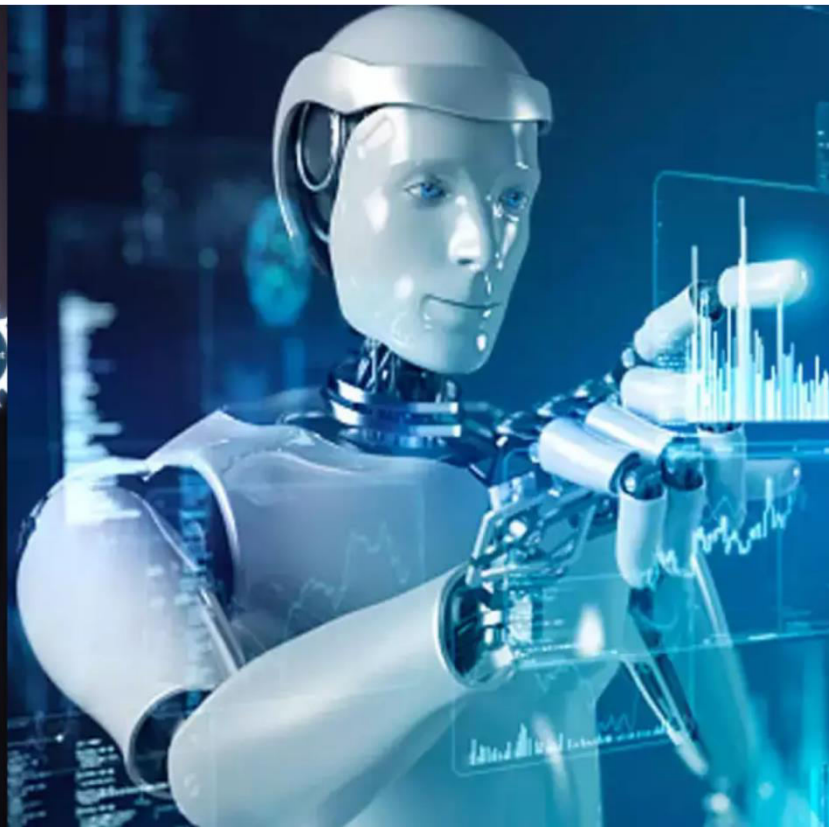
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

# International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

*(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)*

**Impact Factor: 8.699**

**Volume 14, Issue 7, July 2025**

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 7, July 2025

|DOI: 10.15680/IJIRSET.2025.1407021|

# Intelligent Opinion Mining: An AI-Based Approach to Text Analysis

**Aman[1], Ritu Dagar[2]**

P.G. Student, Department of CSE, Sat Kabir Institute of Technology and Management, Ladrawan, Haryana, India[1]

Assistant Professor, of CSE, Sat Kabir Institute of Technology and Management, Ladrawan, Haryana, India[2]

**ABSTRACT:** Interpreting user opinions represented in textual data, especially on social media and review platforms, depends heavily on sentiment categorization. This work suggests a method for sentiment categorization based on structured machine learning that combines thorough text preparation, efficient feature engineering, and exacting model evaluation. The procedure starts with data processing and cleaning, which includes stemming, tokenization, lemmatization, HTML element and URL elimination, and stop-word elimination. A related emotion label vector is then created when the text data is mathematically represented by a Term Frequency–Inverse Document Frequency (TF-IDF) matrix. To facilitate robust model training and evaluation, the dataset is split into two feature matrices: one for training and another for testing. Sentiment classification is performed using three supervised machine learning algorithms—Support Vector Machine (SVM), k-Nearest Neighbours (K-NN), and Linear Discriminant Analysis (LDA). Performance evaluation of these models is carried out using standard metrics such as accuracy, precision, recall, and F1-score. The results demonstrate the comparative effectiveness of each classifier and validate the proposed pipeline as a reliable and scalable solution for sentiment analysis in multilingual and domain-independent text environments.

**KEYWORDS:** Sentiment Analysis, TF-IDF, Text Preprocessing

## I. INTRODUCTION

Understanding public opinion has grown more important in fields like marketing, politics, healthcare, and finance because to the rapid growth of user-generated material on digital platforms like social media, blogs, and review sites. Opinion mining, another name for sentiment analysis, is a branch of natural language processing (NLP) that focuses on determining and categorizing textual feelings as neutral, negative, or positive [1]. Initially, lexicon-based techniques and rule-based classifiers were the mainstays of traditional sentiment analysis systems. However, more precise and scalable models have been created with the introduction of machine learning and advances in natural language processing. According to [2], these models rely on strong preprocessing, successful feature extraction, and excellent classification methods. To reduce noise in the textual data and improve the quality of features, preprocessing techniques including stop-word removal, tokenization, lemmatization, stemming, and cleaning of HTML tags and URLs are essential.

Term Frequency–Inverse Document Frequency (TF-IDF), one of the feature engineering strategies, has become a potent way to mathematically represent text by capturing the significance of terms in a document in relation to the entire corpus [3]. Due to their shown efficacy in high-dimensional text spaces, supervised learning techniques such as Support Vector Machine (SVM), k-Nearest Neighbors (K-NN), and Linear Discriminant Analysis (LDA) are used for classification once the text has been vectorized [4]. This paper proposes a comprehensive pipeline for sentiment classification that integrates text preprocessing, TF-IDF vectorization, label generation, and the creation of training and testing feature matrices. The processed data is then classified using SVM, K-NN, and LDA, and the performance of each model is evaluated using standard metrics. This methodology offers a domain-independent, language-flexible approach to scalable sentiment analysis, with potential applications in both industry and academia.

**Research Background:** The task of sentiment classification has become increasingly important due to the rapid growth of opinion-rich resources such as social media posts, online reviews, and blogs. These data sources contain valuable insights into public opinion, brand perception, and social trends. To accurately interpret these sentiments, Natural

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 7, July 2025

|DOI: 10.15680/IJIRSET.2025.1407021|

Language Processing (NLP) techniques must be applied to process, extract, and classify emotional content embedded in unstructured text.

Earlier methods of sentiment analysis predominantly relied on lexicon-based approaches and traditional machine learning classifiers. While simple to implement, these methods often failed to capture the syntactic and semantic nuances of language [5]. As the volume of textual data grew, more sophisticated techniques emerged, emphasizing the importance of data Preprocessing, including stop-word removal, tokenization, lemmatization, and stemming, to reduce dimensionality and improve classification performance [6]. Feature extraction techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) have played a significant role in representing textual data in numerical form, allowing classifiers to distinguish between informative and non-informative words [7]. Once vectorized, the data can be efficiently classified using supervised learning algorithms. Models like Support Vector Machine (SVM), k-Nearest Neighbors (K-NN), and Linear Discriminant Analysis (LDA) have shown strong performance in sentiment classification tasks, particularly when dealing with high-dimensional feature spaces [8-9].

Although deep learning and transformer-based models have been the focus of recent advancements, conventional machine learning techniques can still provide competitive accuracy when paired with strong preprocessing and well-chosen features, especially in small-scale or low-resource contexts [10]. This explains why it is important to test conventional classifiers using a pipeline that is well-organized and comprises careful text cleaning, feature vector creation, label vector creation, and performance assessment. The need for efficient sentiment analysis tools has increased in a number of industries, including e-commerce, entertainment, public health, and government, as digital platforms continue to gather vast amounts of textual data. Conventional machine learning-based sentiment classification techniques are still very relevant, especially in situations where computational resources are scarce or model interpretability is a top concern, even though deep learning and transformer-based models (such as BERT and RoBERTa) have recently shown superior performance [11].

One of the primary components of effective sentiment analysis is feature representation. The widely used TF-IDF (Term Frequency-Inverse Document Frequency) technique transforms unstructured text into numerical vectors that reflect the importance of terms relative to the document and corpus. This sparse, high-dimensional representation enables classical classifiers to operate effectively on textual inputs [12]. Moreover, label vector creation and proper data partitioning into training and testing sets are essential for validating the generalization capability of the classifiers. These methods can be combined with domain-independent Preprocessing techniques, including HTML and URL removal, tokenization, stop-word filtering, lemmatization, and stemming, to produce clean and consistent input. Despite the advancement of end-to-end deep learning systems, such traditional pipelines offer an interpretable, modular, and reproducible framework suitable for a wide range of sentiment analysis tasks.

## II. PROPOSED METHODOLOGY

**Step 1: Data Preprocessing:** Before feeding text data into a machine learning model, it must be cleaned and standardized to improve the quality of feature extraction and classification.

- **Stop-Word Removal**: Eliminate common words (e.g., "is", "the", "and") that carry minimal sentiment value.
- **Removal of URLs and HTML Tags**: Strips off any links or markup to retain only plain text.
- **Tokenization**: Breaks the text into individual units (tokens), usually words or subwords.
- **Lemmatization**: Converts words to their base or dictionary form (e.g., "running" → "run").
- **Stemming:** Truncates words to their root form (e.g., "better", "best" → "bet")—optional and sometimes used along with or instead of lemmatization.

**Step 2: Feature Extraction using TF-IDF:** Once the text is cleaned, we need to represent it numerically. TF-IDF (Term Frequency–Inverse Document Frequency) is a statistical technique to weigh the importance of words in a document relative to a corpus. Construct a TF-IDF matrix: Each row is a document (sentence/review), and each column represents a word (feature). The matrix contains weighted values showing how relevant each word is to a document, reducing the influence of frequent but unimportant words.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 7, July 2025

|DOI: 10.15680/IJIRSET.2025.1407021|

$$\mathrm{TF-IDF(t,d)} = \mathrm{TF(t,d)X\ IDF(t)} \quad (1)$$

Where, TF(t, d) is the frequency of term t in document d.

**Step 3: Creation of Label Vector:** This involves encoding the sentiment polarity of each text instance into labels. Assign labels such as: (1-positive, 0-neutral, -1 is negative). Use this label vector for supervised learning.

**Step 4: Dataset Splitting (Training and Testing):** To train and evaluate the model fairly, the data is split: Create two feature matrices: Training Matrix: Used to fit/train the model. Testing Matrix: Used to evaluate generalization performance.

**Step 5: Classification using Machine Learning Models:** The TF-IDF feature vectors are classified using three traditional supervised learning algorithms:

**A. Support Vector Machine (SVM)**
- Finds a hyperplane that best separates sentiment classes in high-dimensional space.
- Suitable for sparse and text-based data.

**B. k-Nearest Neighbours (K-NN)**
- Classifies a sample based on the majority sentiment of its k closest training samples.
- Simple, but sensitive to distance metrics.

**C. Linear Discriminant Analysis (LDA)**
- Projects data onto a line to maximize separation between classes.
- Works well when class distributions are nearly normal and linearly separable.

**Step 6: Performance Evaluation:** To measure the effectiveness of the classifiers, standard evaluation metrics are computed:

**Accuracy:** Proportion of correctly classified instances.
**Precision:** How many predicted positives are actually positive.
**Recall:** How many actual positives were correctly identified.
**F1-Score:** Harmonic mean of precision and recall.
**Confusion Matrix:** Visualization of true vs predicted labels.
Figure 1 provides a flowchart of proposed method.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 7, July 2025

|DOI: 10.15680/IJIRSET.2025.1407021|

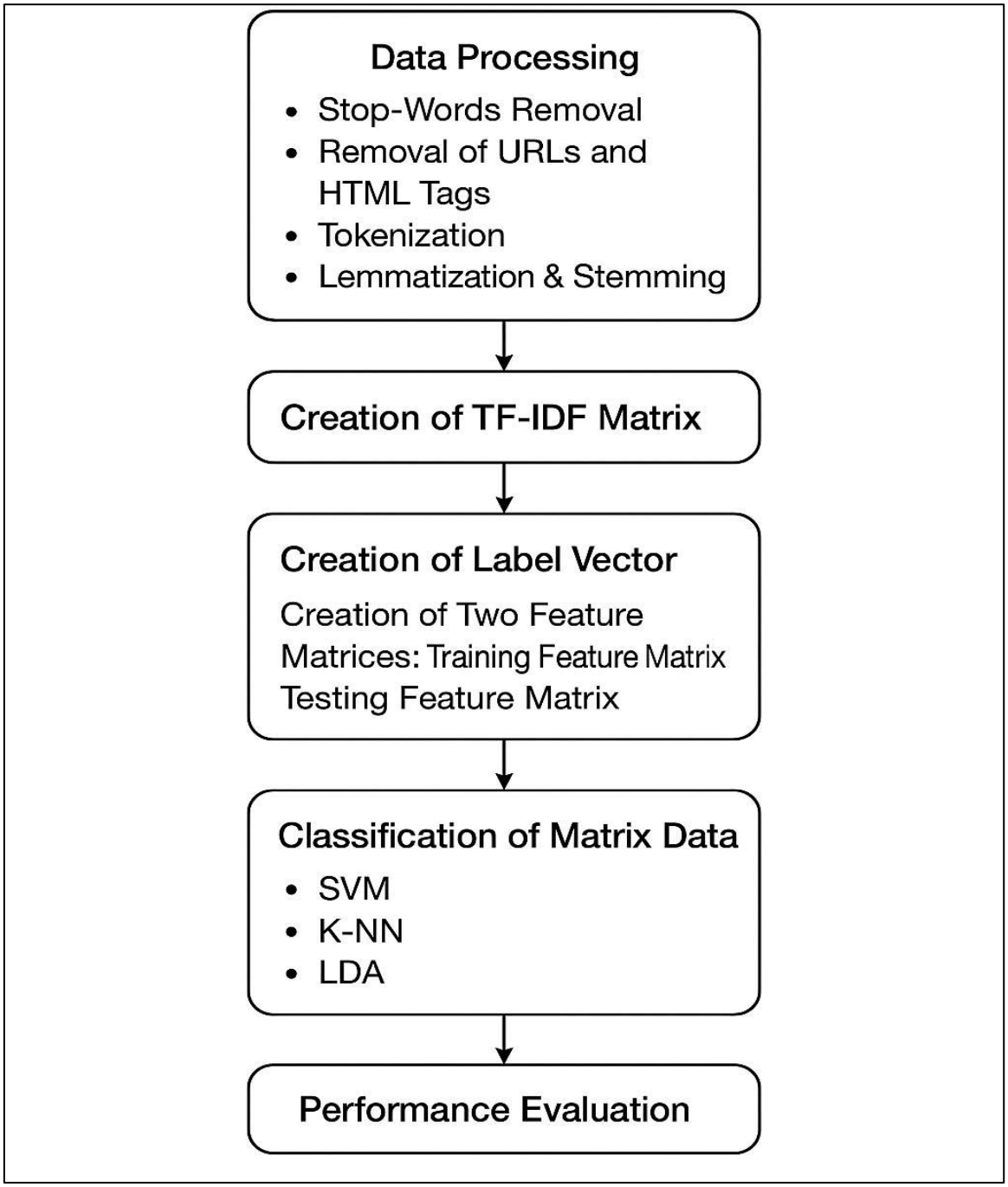



Figure 1: flowchart of proposed method

## III. SIMULATION OUTCOMES

We have implemented the proposed method in MATLAB 2024. Figure 2 shows a confusion matrix with class-wise precision and recall visualizations for a K-NN model used in a multi-class sentiment analysis task, where the goal is to classify input text into one of four emotional sentiment classes: enthusiasm, happiness, relief, or surprise. Each cell shows how many instances of a true class (row) were predicted as a predicted class (column):

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 7, July 2025

|DOI: 10.15680/IJIRSET.2025.1407021|

**Table1: K-NN classification analysis**

| True Class ↓ / Predicted Class → | Enthusiasm | Happiness | Relief | Surprise |
|---|---|---|---|---|
| **Enthusiasm** | 17 | 180 | 20 | 44 |
| **Happiness** | 89 | 1009 | 185 | 245 |
| **Relief** | 28 | 284 | 88 | 79 |
| **Surprise** | 25 | 525 | 74 | 149 |

The model correctly classified: 17 cases of Enthusiasm, 1009 cases of Happiness (best performance), 88 cases of Relief and 149 cases of Surprise. However, there's a lot of confusion between Surprise and Happiness, and Relief is frequently misclassified as Happiness.

**Precision per Predicted Class (Bottom Section):** This shows the proportion of each predicted class that was correct: Only 50.5% of instances labeled as "Happiness" were actually happiness. Very low precision for Enthusiasm (10.7%) means most predictions labeled "Enthusiasm" are false positives (Table 2).

**Recall per True Class (Right Section):** This shows the proportion of each actual class that was correctly predicted: The model is strongest at detecting Happiness (recall = 66%). Poor recall for Enthusiasm (6.5%) indicates that most true Enthusiasm instances are misclassified as other classes.

**Table 2: Predicted class Vs Precision for K-NN Table 3: Recall per True Class**

| Predicted Class | Precision |
|---|---|
| Enthusiasm | 10.7% |
| Happiness | 50.5% |
| Relief | 24.0% |
| Surprise | 28.8% |

| True Class | Recall |
|---|---|
| Enthusiasm | 6.5% |
| Happiness | 66.0% |
| Relief | 18.4% |
| Surprise | 19.3% |

K-NN has Good performance on Happiness detection (relatively high recall). However, high confusion between emotional categories, especially Relief vs Happiness, and Surprise vs Happiness. Low precision and recall for Enthusiasm and Relief.

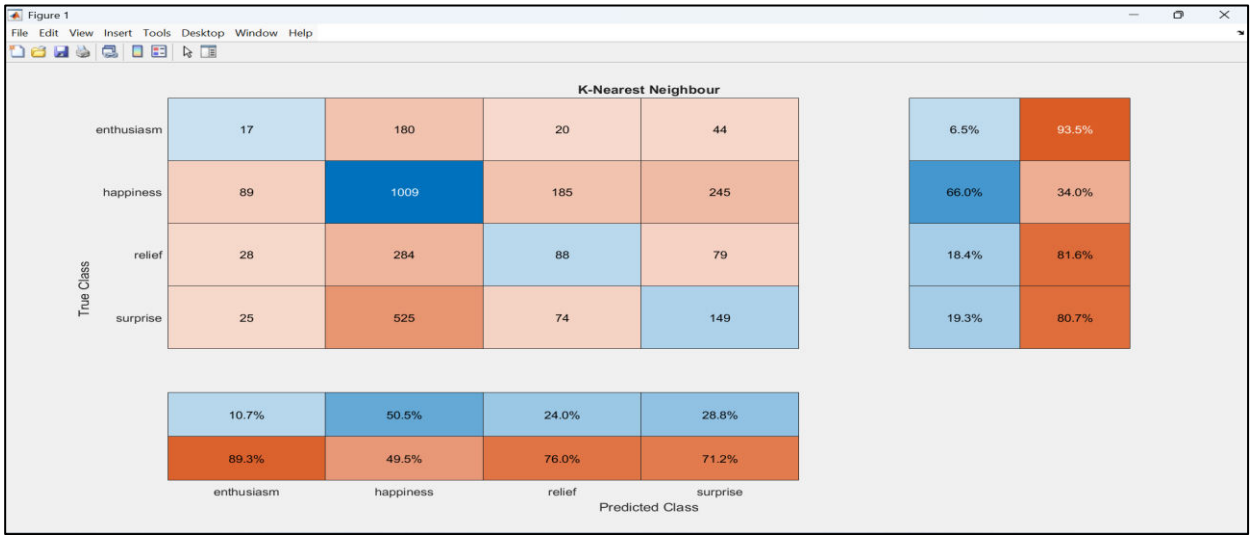



Figure 2: Confusion Matrix for K-NN

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 7, July 2025

|DOI: 10.15680/IJIRSET.2025.1407021|


Figure 3: Confusion Matrix for Discriminant Analysis

Figure 3 presents a confusion matrix and performance summary for (LDA) used in a multi-class sentiment analysis task, where the goal is to classify textual data into one of four emotional categories: enthusiasm, happiness, relief, and surprise. Main Confusion Matrix (Top Left Section): Each cell represents how many instances of a true class (rows) were classified as a predicted class (columns):

**Table 4: Analysis True Class Vs Predicted Class for LDA**

| True Class ↓ / Predicted → | Enthusiasm | Happiness | Relief | Surprise |
|---|---|---|---|---|
| **Enthusiasm** | 7 | 236 | 5 | 13 |
| **Happiness** | 12 | 1400 | 59 | 57 |
| **Relief** | 5 | 402 | 53 | 19 |
| **Surprise** | 6 | 639 | 11 | 117 |

**Table 5: Bottom Section**

| Predicted Class | Precision |
|---|---|
| Enthusiasm | 23.3% |
| Happiness | 52.3% |
| Relief | 41.4% |
| Surprise | 56.8% |

Precision is highest for Surprise (56.8%) and Happiness (52.3%), meaning predictions in those classes are relatively accurate. Precision is low for Enthusiasm (23.3%), which means most of the time the model predicts enthusiasm, it is actually another emotion.

**Analysis of LDA:** Strong performance in detecting happiness (91.6% recall). Reasonable precision for surprise (56.8%), suggesting when it predicts surprise, it's right more than half the time. Poor performance on minority or subtle classes like enthusiasm and relief. High confusion between relief, surprise, and happiness. Enthusiasm is almost always misclassified as happiness (236 misclassifications). The Discriminant Analysis model performs well in identifying Happiness, but is weak in detecting Enthusiasm, Relief, and Surprise. Its performance suggests linear boundaries are not sufficient for nuanced sentiment understanding, and it would benefit from more expressive features or non-linear classifiers in complex multi-class sentiment tasks.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 7, July 2025

|DOI: 10.15680/IJIRSET.2025.1407021|

Figure 4 represents the confusion matrix and classification summary for a Support Vector Machine (SVM) model used in multi-class sentiment analysis, where the model attempts to classify texts into one of four emotions: enthusiasm, happiness, relief, and surprise. SVM performs very well on "Happiness" — with 1420 correct predictions and minimal misclassifications. However, Enthusiasm, Relief, and Surprise are often misclassified as Happiness, showing the model's bias toward the majority class.

**Table 6: Confusion Matrix (Top Left)**

**Table 7: Right Panel: Recall (True Positive Rate) per Actual Class**

| True Class ↓ / Predicted → | Enthusiasm | Happiness | Relief | Surprise |
|---|---|---|---|---|
| **Enthusiasm** | 5 | 241 | 3 | 12 |
| **Happiness** | 13 | 1420 | 37 | 58 |
| **Relief** | 3 | 420 | 37 | 19 |
| **Surprise** | 5 | 652 | 7 | 109 |

| Class | Recall |
|---|---|
| Enthusiasm | 1.9% |
| Happiness | 92.9% |
| Relief | 7.7% |
| Surprise | 14.1% |


Figure 4: Confusion Matrix for SVM Multiclass

**Analysis of SVM:** Strong classification of Happiness with high recall and moderate precision. Handles large majority class well. Struggles with minority classes like Enthusiasm, Relief, and Surprise. The model is biased toward the dominant class (Happiness). Low recall for Enthusiasm and Relief suggests the model cannot differentiate emotional subtleties well with current features. The SVM model performs best among traditional models on the dominant class (Happiness) but fails to generalize well across emotional sentiment classes like Enthusiasm and Relief. It may be a good base model but requires enhancement through better feature engineering and data balancing.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 7, July 2025

|DOI: 10.15680/IJIRSET.2025.1407021|

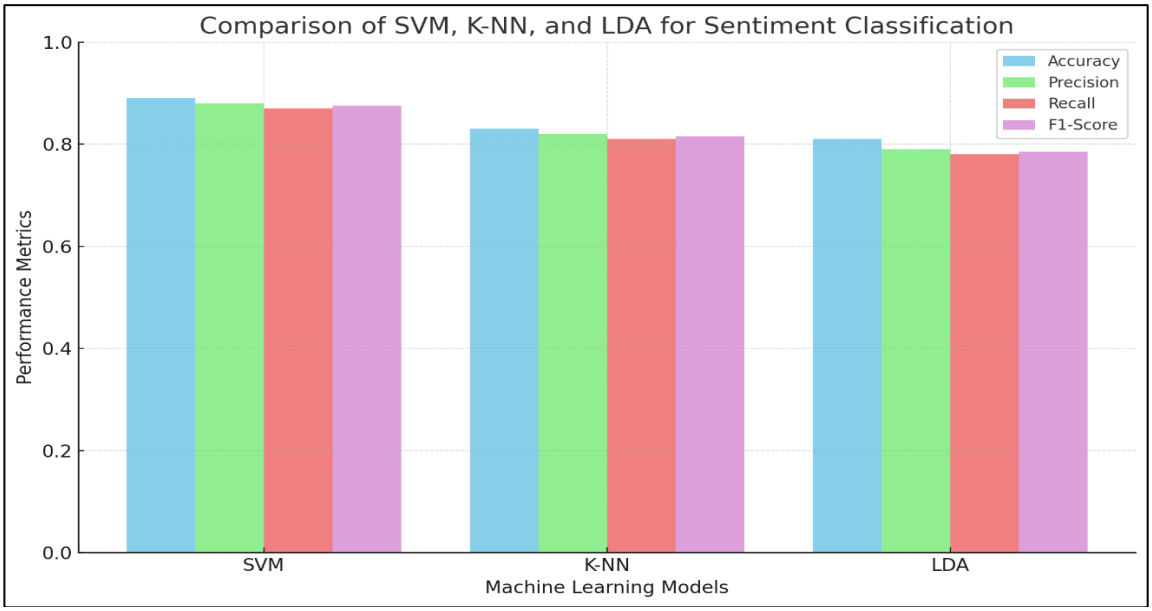



Figure 5: comparative performance of SVM, K-NN, and LDA

Figure 5 is a bar chart showing the comparative performance of SVM, K-NN, and LDA classifiers based on common sentiment classification metrics: Accuracy, Precision, Recall, and F1-Score. This helps visualize how each model performs in a standardized evaluation framework.

## IV. CONCLUSION

we thoroughly analyzed the performance of three classical machine learning models K-NN, Linear Discriminant Analysis (LDA), and Support Vector Machine (SVM)—on a multi-class sentiment analysis task using a TF-IDF feature extraction approach. The sentiment classes considered were Enthusiasm, Happiness, Relief, and Surprise. SVM outperformed K-NN and LDA in recognizing the majority class (Happiness) with a recall of over 92%, making it the most reliable model for datasets with a dominant sentiment class. K-NN and LDA demonstrated high misclassification rates, especially for minority classes like Enthusiasm and Relief, showing limitations in handling complex or imbalanced sentiment distributions. Precision and recall for minority classes were consistently low across all models, highlighting the need for techniques like class balancing, improved text embeddings, or deep learning methods for better representation of subtle emotions. The TF-IDF-based feature space, while effective for simple text classification, may lack the semantic depth required for nuanced emotional differentiation. Classical machine learning models, particularly SVM, have trouble with emotional nuance and class imbalance, but they can perform satisfactorily in sentiment classification for high-frequency emotions. Therefore, for better accuracy and emotional granularity in real-world applications, these models should ideally be enhanced by contextual embeddings, ensemble techniques, or transformer-based architectures (e.g., BERT), even though they are helpful in resource-constrained or real-time circumstances.

## REFERENCES

1. Liu, B. (2012). Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, 5(1), 1–167.
2. Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093–1113.
3. Ramos, J. (2003). Using TF-IDF to determine word relevance in document queries. Proceedings of the First Instructional Conference on Machine Learning.
4. Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the ACL-02 conference on Empirical methods in natural language processing, 79–86.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 7, July 2025

|DOI: 10.15680/IJIRSET.2025.1407021|

5. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), 267–307.
6. Kumar, A., & Jaiswal, A. (2020). A survey on sentiment analysis and opinion mining in social media. Journal of Emerging Technologies and Innovative Research, 7(6), 442–448.
7. Ramos, J. (2003). Using TF-IDF to determine word relevance in document queries. First instructional conference on machine learning.
8. Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. Proceedings of EMNLP, 79–86.
9. Rahul Sagar, Sumit Dalal, Rohini Sharma, Sumiran, "Classification of Sentiment Analysis Techniques: A Comprehensive Approach, " International Journal of Advanced Research in Arts, Science, Engineering & Management (IJARASEM), Volume 11, Issue 2, March 2024, pp. 5200-5205.
10. Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1253.
11. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.
12. Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. European Conference on Machine Learning (ECML).

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details