International Journal of Advance and Applied Research
www.ijaar.co.in
ISSN – 2347-7075 Impact Factor – 8.141
Peer Reviewed Bi-Monthly
Vol. 6 No. 22 March - April - 2025

# Machine Learning Empowered Sentiment Analysis: Techniques and Insights

**Mrs. Ashwini J. Shinde**
*Research Scholar, Department of Computer Science PVG's College of Science and Commerce, Pune-09.*
***Corresponding Author – Mrs. Ashwini J. Shinde***
**DOI - 10.5281/zenodo.15542707**

***Abstract:***

*Sentiment analysis is a crucial area of contemporary research, particularly useful for examining text data and identifying sentiment elements. E-commerce platforms generate vast amounts of text each day through customer comments, reviews, tweets, and feedback. The rise of social networking sites has significantly enhanced communication and knowledge-sharing. Conducting aspect-based sentiment evaluations can provide businesses with valuable insights into consumer expectations, enabling them to adjust their strategies accordingly. Conveying the precise sentiment of a review can be challenging. This study introduces an approach focusing on the sentimental aspects of product characteristics. We analysed consumer reviews from Amazon and IMDB, using a dataset sourced from the UCI repository, which includes opinion ratings for each review. To derive meaningful information from the datasets and reduce noise, we performed pre-processing steps such as tokenization, punctuation removal, and elimination of whitespace, special characters, and stop words. For effective representation of the pre-processed data, we applied feature selection methods like term frequency-inverse document frequency (TF-IDF). We merged customer reviews from three datasets—Amazon, Yelp, and IMDB—before applying classification using algorithms including Naïve Bayes, Random Forest, K-Nearest Neighbour (KNN), and Support Vector Machine (SVM). Finally, we offer insights into potential future work in text classification.*

***Keywords- sentimental analysis, TF-IDF, Amazon, IMDB, KNN, SVM, Naïve Bayes, yelp, feature extraction.***

## Introduction:

Sentiment analysis is a prominent machine learning technique that helps identify emotions, enabling entrepreneurs to gain insights into customer opinions through various online platforms, including surveys, social media, and e-commerce reviews. With the global commercial sector largely shifting to online markets, consumers often review products or services before making purchases. Consequently, analysing customer feedback has become increasingly vital for businesses, as modern consumers frequently rely on reviews for making decisions about products and movies.

Applications of sentiment analysis in industry range from predicting future market trends based on sentiments expressed in blogs and news articles to assessing customer satisfaction and dissatisfaction through social media posts and product reviews. This analysis also lays the groundwork for various other applications, such as recommender systems. The complexity of sentiment analysis entails removing noisy data from the original datasets, selecting relevant features, and choosing appropriate classifiers. The field gained traction in the early 2000s, attracting significant scholarly interest. Two key

concepts in sentiment analysis are 'Polarity' and 'Subjectivity.' Subjectivity refers to individual beliefs, opinions, or personal sentiments, while polarity reflects feelings expressed as negative, positive, or neutral. Sentiment analysis can be conducted at different levels, including sentence, sub-sentence, and document levels. Various forms include fine-grained sentiment analysis, which operates on polarities from very negative to very positive, as well as intent-based or emotion detection analyses. Traditional lexicon-based methods and machine learning techniques can be employed for sentence analysis, each with advantages and limitations. This study aims to identify and classify positive and negative customer feedback on various products and film reviews using a machine learning model. An analysis conducted on Amazon last year revealed that over 80% of online customer's valued reviews more than specific recommendations. A product with a high number of positive reviews significantly enhances its credibility, while a lack of reviews can leave potential customers skeptical. Essentially, an abundance of reviews tends to correlate with greater credibility. Consumers often rely on others' opinions and experiences to gauge product quality, and reviews play a crucial role in this process. Negative reviews can lead to a decline in sales. In this study, we merged two datasets from Amazon and IMDB for movie reviews. After pre-processing the data, we performed feature extraction using the TF-IDF approach and applied machine learning classifiers to evaluate performance results.

The remainder of the paper is structured as follows: Section 2 provides a brief overview of the literature on sentiment analysis. Section 3 outlines the methodologies employed. Section 4 presents a comprehensive experimental analysis, and Section 5 concludes with a discussion on future work.

**Related Work:**

A significant amount of research has been conducted in the domains of text classification and sentiment analysis, with a fundamental challenge being the categorization of sentiment scores. This task involves determining whether a segment of text conveys a negative or positive sentiment. Sentiment polarity classification can take three distinct forms: aspect or entity level, document level, and sentence level. The entity level focuses on individuals' opinions regarding their preferences, while the document level assesses the overall valence (positive or negative) of the complete text. Sentence-level analysis, on the other hand, evaluates sentiment categorization on a per-sentence basis.

Researchers worldwide have explored various methodologies using semi-supervised, unsupervised, and supervised machine learning techniques. For example, Bhatt et al. proposed a sentiment analysis methodology to evaluate iPhone 5 reviews on Amazon. Their approach utilized various preprocessing techniques to reduce noise from the data, including the removal of punctuation, numbers, and HTML tags. They employed part-of-speech (POS) tagging for feature identification and a rule-based method for categorizing reviews. Shrestha and Nasoz investigated the consistency between Amazon.com reviews and corresponding ratings, as discrepancies sometimes arise between user-submitted reviews and ratings. They employed deep learning techniques for sentiment analysis on product reviews from Amazon.com to identify mismatched ratings. By converting product reviews into paragraph vectors, they trained a recurrent neural network (RNN) using gated recurrent units, considering both product information and the semantic relationships within review text. Their trained model created a web service that predicts the rating score based on submitted

reviews and provides feedback when discrepancies between submitted and predicted scores occur. Haddi et al. examined the impact of text preprocessing on online movie reviews, implementing techniques like stemming, HTML tag removal, and data cleaning to eliminate noisy data. They applied the chi-square method for feature selection, removing irrelevant features, and employed Support Vector Machine (SVM) to classify reviews into categories of negative or positive sentiment. Chen et al. conducted review embedding on the IMDB and Yelp datasets using a one-layer convolutional neural network (CNN). This CNN produced 300-dimensional vectors from reviews of varying lengths by padding shorter reviews with zero vectors for uniformity. They applied filters of widths 3 and 5 for one-dimensional convolution on the word embeddings, generating multiple feature maps. The max-pooling layer then captured only the most relevant features. These outputs were concatenated to create a 300-dimensional vector. To train the network across K classes, the Softmax function was employed as the activation function, ensuring effective classification. Wassan et al. introduced a novel approach aimed at the sentimental aspects of item characteristics. They focused on customer reviews from Amazon, deriving datasets from Data World Center to assess opinion rates and extract significant information related to negativity or positivity. Their preprocessing steps included tokenization, stemming, removal of stop words, and boxing, with a focus on examining data at the aspect level to understand consumer preferences and guide future behavior. Srujan et al. conducted noise elimination from Amazon book reviews using various preprocessing techniques, such as URL and HTML tag removal, whitespace and punctuation cleaning, stemming, and special character removal. They utilized Term Frequency-Inverse Document Frequency (TF-IDF) for feature selection and evaluated the accuracy and processing time of classifiers including K-Nearest Neighbor (KNN), Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), and Naive Bayes (NB), alongside determining sentiment scores for different books.Nandal et al. proposed an innovative method for aspect-level sentiment detection that emphasizes item features. They tested their approach on Amazon customer reviews by first identifying aspect terms in each review. Preprocessing operations, such as tokenization, stemming, stop-word removal, and casing, were conducted to derive meaningful insights before classifying data as positive or negative.

Following this, they utilized a Recurrent Neural Network (RNN) to capture temporal information and integrate both user and product input. Their findings reported state-of-the-art results on the Yelp and IMDB datasets, based on the premise that items receiving positive feedback initially are likely to receive similar feedback in the future, and vice versa. In my work, we present a methodology focused on noise elimination and feature extraction using TF-IDF from Amazon and IMDB datasets. We combine these datasets to perform a comparative analysis of the accuracy of various classifiers and their sentiment scores. Details regarding the methodology and classifiers used in our experiment will be discussed in the subsequent section.

**Methodology for the Sentimental Analysis:**

In our analysis, we will leverage the Amazon product review, IMDB movie review, and Yelp review datasets. The methods employed include preprocessing, classification, and representation techniques. Preprocessing entails data cleaning, removal of numbers and punctuation, elimination of stop words, and stripping of HTML/URL

tags. After preprocessing, we use the TF-IDF representation model to convert the text into a usable format. To classify the dataset into negative and positive categories, we will implement classifiers such as Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbours (KNN). The performance of these classifiers will be compared. The overall workflow is illustrated in Figure 1.

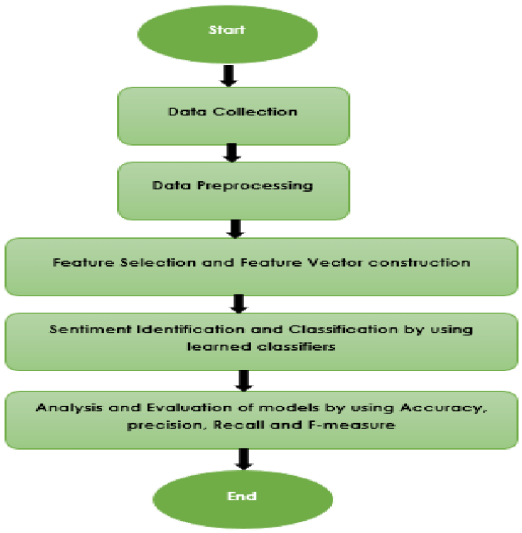


Fig. 1 Workflow of Sentiment Analysis

**Data Acquisition:**

We acquired three datasets: Amazon product reviews, IMDB movie reviews, and Yelp reviews from the UCI repository. These reviews were merged into a single dataset. Since all reviews are rated on a 5-star scale, we classified 3-star ratings as neutral, indicating that they are neither negative nor positive. Consequently, we removed all 3-star reviews from the dataset. In this classification, a rating of 1 is designated as positive, while a rating of 0 is treated as negative.

**Data Processing:**

Data preprocessing refers to the process of transforming raw data into a structured format to enhance data quality. This method encompasses a variety of operations, including tokenization, punctuation removal, whitespace, special character elimination, and the removal of stop words.

**Tokenization:**

Tokenization involves breaking down patterns into distinct components such as phrases, words, symbols, keywords, and other defined elements known as tokens. Tokens can consist of phrases, individual words, or complete sentences. During tokenization, certain characters, particularly punctuation marks, are typically discarded. These tokens serve as input for processes like text mining and parsing. Eliminating stop words is another critical step in data preprocessing. Stop words are elements within a sentence that do not contribute meaningful information for text mining. To improve the accuracy of analyses, these words are usually excluded. The list of stop words can vary by language and country. In English, common stop words include 'or', 'each', 'both', 'when', and 'whereupon', among others. Parts of Speech (POS) tagging, commonly referred to as POS tagging, is the process of assigning a specific part of speech to each word in a sentence. The categories include verbs, adverbs, nouns, pronouns, adjectives, conjunctions, and their respective subclasses. A program designed to carry out this task is known as a Parts of Speech tagger, or POS tagger.

**Feature Extraction:**

A crucial step in sentiment classification is data representation. Raw data often contains noise that needs to be filtered out using various preprocessing methods. Once preprocessed, the data is transformed into a term-document matrix (TDM), which calculates the frequency of each word. Two prominent feature extraction methods supported by the TDM are TF-IDF and the bag of words model. The TF-IDF score for a word is determined by the product of its Term Frequency (TF) and Inverse Document Frequency (IDF). In this scoring system, words that appear frequently

in the review dataset are assigned higher TF values, while the IDF scaling factor gives more importance to less common words. Consequently, common words receive lower scores. By ignoring terms with low TF-IDF values, we can effectively eliminate less significant words from the dataset**.**

TF-IDF is a method used for information extraction that considers both Term Frequency (TF) and Inverse Document Frequency (IDF). Each word or phrase is assigned a distinct TF score and an IDF score. The TF*IDF weight of a term is calculated by multiplying its TF and IDF scores. In essence, a higher TF*IDF score indicates that the term is relatively rare, while a lower score suggests otherwise. The TF of a term quantifies how often it appears in the text, whereas the IDF reflects its overall importance across the entire dataset.

When words in content have a high TF*IDF weight, the information will likely rank among the top search results, enabling us to:

1. Stop worrying about the inclusion of stop words.
2. Focus on identifying words that have low competition but high search volume.

**Experimentation:**

After transforming the training and testing datasets into separate term-document matrices (TDM), the occurrence frequency of each word is analyzed. These TDMs are subsequently inputted into classifiers such as SVM, KNN, Naive Bayes, and Random Forest. The algorithm for the proposed work is outlined below.

Algorithm:

Input: Labeled data

Output: Classifier accuracy; Recall, Precision, F-1 Measure for positive and negative values

1. Load labeled data for negative (0) and positive (1) classes.
2. Preprocess the labeled data.
3. For each value {V1...Vn} in the labeled dataset:
4. Perform feature extraction for each feature (Vi).
5. Validate the data by splitting it into training and testing sets.
6. Train the classifier.
7. Calculate the accuracy of the model.
8. Compare the results (precision, recall, accuracy, F-1 Measure) across different machine learning classifiers, such as SVM, KNN, Naive Bayes, and Random Forest.
9. End.

**Results and Discussion:**

This section provides a comprehensive overview of the evaluation metrics, along with a discussion of the results obtained.

**Evaluation Metrics:** Evaluation metrics are essential for assessing the performance of classification models. Accuracy is the most commonly used metric for this purpose; it represents the proportion of correctly classified instances in a specified test dataset. However, accuracy alone is insufficient for making robust judgments in text mining analysis. Therefore, we employ additional metrics to evaluate classification performance. Three fundamental measures are typically used: F-measure, recall, and precision. Before exploring these measures further, it's important to familiarize ourselves with a few key terms:

- TP (True Positive): The count of data points that were correctly classified as positive.
- FP (False Positive): The count of incorrectly classified positive instances.
- FN (False Negative): The count of actual positive instances that were incorrectly classified as negative.
- TN (True Negative): The count of actual negative instances that were correctly classified as negative.

- **Accuracy**: Accuracy reflects how frequently the classifier makes correct predictions. It is calculated by dividing the number of correct predictions by the total number of predictions made.

$Accuracy$ = Correct Prediction /Total prediction

**Recall**: Recall measures a classifier's sensitivity, with higher recall indicating fewer false negatives. It is calculated as the number of correctly classified positive instances divided by the total number of actual positive occurrences. This can be represented mathematically as:

$Recall$ (R) = TP / (TP + FN)

**Precision**: Precision assesses the accuracy of a classifier's positive predictions. A low precision indicates a higher number of false positives, while a high precision signifies fewer such errors. Precision (P) is defined as:

$Precision$ $(P)$ = TP / (TP + FP)

**F-Measure:** The F-measure combines precision and recall into a single metric, known as the weighted harmonic mean of the two. It is defined as:

$F-measure$ = 2PR/P+R

**Experimental Results:**

In this study, we analyze two classes of sentiments: negative and positive. We collected reviews from three different platforms: Amazon, IMDB, and Yelp. Each review includes a user's rating and text feedback. Reviews rated 4 or 5 are classified as positive, while those rated 1 or 2 are deemed negative. The dataset is split into training and testing sets in an 80:20 ratio.

We employ various machine learning algorithms for classification, including Support Vector Machine Classifier (SVC), K Nearest Neighbor (KNN), Naive Bayes, and Random Forest. Several preprocessing steps are performed on the dataset, such as data cleaning, removal of extra whitespace, and stemming. Both the training and test sets are converted into individual term document matrices (TDM), where the frequency of each word is counted. The TF-IDF method is then applied to represent the data. Following this, the TDM serves as input for multiple classifiers.

The experimental results for different classifiers are displayed in Table 1. From the data, it is evident that the Random Forest outperforms the other methods in the two-class classification task, achieving the highest accuracy of 78.96%. The Term Document Matrix (TDM) consists of feature vector values referred to as points. The classification accuracy was significantly enhanced by employing an ensemble of trees that collectively vote on the most prevalent class.

| Dataset | Classifier | Accuracy | Precision | Recall | F1 score |
|---|---|---|---|---|---|
| AMAZON+IMDB+YELP | support Vector machine | 0.76 | 0.76 | 0.76 | 0.76 |
| | Random Forest | 0.78 | 0.79 | 0.78 | 0.78 |
| | Multinomial Naïve Bayes | 0.77 | 0.77 | 0.77 | 0.77 |
| | KNN | 0.61 | 0.62 | 0.61 | 0.61 |

**Conclusion:**

Customer sentiment analysis is crucial for online retailers to understand their customers' responses effectively. By analyzing reviews, retailers can create recommendation systems tailored to individual preferences. This study explored various machine learning models on a large, unlabeled dataset of reviews from Amazon, IMDb, and Yelp. We employed different

feature extraction techniques and outlined the theoretical underpinnings of each model, alongside the methodologies used in our research and performance metrics from extensive experiments. The Random Forest classifier achieved an accuracy exceeding 78%. To enhance representation, we suggest expanding the research scope by incorporating feature selection techniques like mutual information (MI), information gain, and the chi-squared test. Additionally, combining hybrid classifiers, such as SVM with other algorithms, can further improve accuracy. A robust recommendation system can be developed by considering the emotions reflected in customer reviews.

**References:**


1. Gupta, K., Jiwani, N., Whig, P. (2023). Effectiveness of machine learning in detecting early-stage leukemia. International Conference on Innovative Computing and Communications.
2. Miao, Q., Li, Q., Dai, R. (2009). AMAZING: A sentiment mining and retrieval system. Expert Systems with Applications.
3. Srujan, K.S, Nikhil, S.S, Rao, R., Kedage, K., Harish, B.S., Kumar, H.M.K. (2018). Classification of Amazon Book Reviews Based on Sentiment Analysis. Information Systems Design and Intelligent Applications.
4. Haddi, E., Liu, X., Shi, Y. (2013). The role of text preprocessing in sentiment analysis. Procedia Computer Science.
5. Prabowo, R., Thelwall, M. (2009). Sentiment analysis: A combined approach. Journal of Informetrics.
6. Nandal, N., Tanwar, R., Pruthi, J. (2020). Machine learning based aspect level sentiment analysis for Amazon products. Spatial Information Research.
7. Alsaeedi, A., Khan, M.Z. (2019). A study on sentiment analysis techniques of Twitter data. International Journal of Advanced Computer Science and Applications.
8. Wassan, S., Chen, X., Shen, T., Waqar, M., Jhanjhi, N.Z. (2021). Amazon product sentiment analysis using machine learning techniques. Revista Argentina de Clínica Psicológica.
9. Bhatt, A., Patel, A., Chheda, H., Gawande, K. (2015). Amazon review classification and sentiment analysis. International Journal of Computer Science and Information Technologies.
10. Vijayarani, S., Ilamathi, M.J., Nithya, M. (2015). Preprocessing techniques for text mining-an overview. International Journal of Computer Science & Communication Networks.
11. Shrestha, N., Nasoz, F. (2019). Deep learning sentiment analysis of amazon.com reviews and ratings.
12. Jiwani, N., Gupta, K., Whig, P. (2023). Assessing permeability prediction of BBB in the central nervous system using ML. International Conference on Innovative Computing and Communications. Lecture Notes in Networks and Systems, Springer, Singapore.
13. Tan, W., Wang, X., Xu, X. (2018). Sentiment analysis for Amazon reviews. In International Conference.
14. Chen, W., Lin, C., Tai, Y.S. (2015). Text-based rating predictions on Amazon health and personal care product reviews. Computer Science.
15. Wang, M., Qiu, R. (2015). Text mining for Yelp dataset challenge. Computer Science.