

2505091 - TEJA VAMSI DEVALLA

by Teja Vamsi Devalla

Submission date: 16-Apr-2026 08:44PM (UTC+0100)

Submission ID: 279995752

File name: 65886_Teja_Vamsi_Devalla_2505091_-_TEJA_VAMSI_DEVALLA_726057_873428976.docx (61.51K)

Word count: 3279

Character count: 20176

TRADITIONAL MACHINE LEARNING APPROACH FOR SENTIMENT ANALYSIS

Module Name: Application Of Machine Learning

Student Name: Teja Vamsi Devalla

Student ID: 2505091

ABSTRACT:

Sentiment analysis completely focuses on identifying and classifying opinions expressed in textual data such as customer reviews. With the rapid growth of unstructured text generated in digital platforms, this highlights the need for efficient methods to automatically analyse and interpret this data to support decision-making. This study selects machine learning algorithm for sentiment analysis to the customer review. In this study, several traditional machine learning algorithms, including Naïve Bayes, Logistic Regression, Random Forest, K-Nearest Neighbour and Support Vector Machine (SVM), were considered along with standard preprocessing and feature extraction techniques. The models were evaluated based on their performance in sentiment classification tasks [1]. Based on the literature, SVM along with feature extraction TF-IDF is recommended as the most suitable method for marketing company sentiment analysis on customer review. This setup provides strong accuracy, handles high-dimensional text data with high efficiency, and sustains a balance between performance and computational cost. Previous studies also show the effectiveness of SVM in sentiment analysis tasks [1]. Overall, this study shows that a simple, efficient, and traditional machine learning algorithm can produce reliable sentiment analysis outcomes suitable for real-world applications.

INTRODUCTION:

Sentiment analysis the way of analysing the human behaviour like emotions, opinion and feeling for the subject which is in the form of text. Companies do sentiment analysis to analyse opinions of their customers [1]. By analysing the customers opinions, companies take actions according to their customer's sentiment. A marketing company can do sentiment analysis of shopper reviews by using traditional machine learning approach to minimise the cost and time and the marketing company can run their workflow with basic resources. To know the sentiment of the shopper reviews, suggesting a traditional machine learning approach which can achieve accuracy and efficiency. To propose the best machine learning approach and workflow reviewed the literature of sentiment analysis to compare ML algorithms then design the workflow and recommend the suitable model to the marketing company. In related work, the literature of sentiment analysis by different machine learning approaches are discussed, workflow methodology

explains about the pipeline of sentiment analysis and in discussions and recommendations section suggests the suitable approach and model for the marketing company to do sentiment analysis.

RELATED WORK:

The process of finding emotions in text to understand people's opinion is sentiment analysis, the sentiment will be classified as positive, neutral or negative. Companies like Amazon, Netflix, social media platforms, governments, political parties, movie industries, marketing teams and anyone who wants to understand people's opinions will use sentiment analysis [2]. Many companies collect large amount of text data like reviews, feedback, survey responses for their future growth. Previously these marketing companies' manual read text data for the sentiment analysis which is slow and costly. Researchers found traditional machine learning methods to classify the sentiment into positive, negative or neutral category.

Traditional machine learning is highly useful for sentiment analysis, particularly when companies want models that are simpler, cheaper and easier to run their workflow on less resources than deep learning systems. The workflow in the literature is collecting data, data preprocessing, feature extraction, model training and evaluation. The preprocessing removes noise like unnecessary symbols, punctuations, and stop words, while feature extraction methods such as TF-IDF, Bag of Words (BoW) and n-grams convert text into numbers because machine learning algorithms need numerical data not text.

A comparison study shows that traditional machine learning algorithms are frequently addressed in the literature, including Support Vector Machine (SVM), Naïve Bayes, Logistic Regressing, Random Forest, Decision Tree and k-Nearest Neighbours. Among these algorithms, SVM seems most frequently as a strong algorithm for text classification tasks. In a comparative study of different machine learning algorithms for sentiment analysis, SVM achieved the best accuracy of 82% on the dataset used in that study, SVM algorithm outperformed Naïve Bayes, Decision Tree and Random Forest [1]. And another study comparing SVM and Naïve Bayes on review data also reported that SVM provided the best accuracy, showing its strength in identifying positive and negative sentiment from text [3].

The benefits of SVM are also mentioned clearly in recent study. SVM outperforms even on small datasets and that a kernel is most suitable for text classification because it is computationally simple and effective for separating classes in structured text data [4]. This is crucial for marketing companies because many firms may not have large computing resources or very large, labelled datasets. In this situation, SVM gives a practical balance between accuracy and efficiency.

Logistic Regression is another strong traditional model addressed in the literature. A study comparing Logistic Regression, SVM and Bernoulli Naïve Bayes for sentiment analysis, logistic Regression reached the highest accuracy at 83%, somewhat above SVM

at 81% and Bernoulli Naïve Bayes at 80% [5]. Another comparative text classification study showed that Logistic Regression with bi-grams obtained the highest accuracy in as 80:20 split, while SVM performed best with TF-IDF [6]. These results suggest that Logistic Regression can also be a strong choice, particularly when feature engineering is done well.

Naïve bayes is widely recognised in the literature as one of the simplest and fastest models for text classification. Researchers often describe it as computationally efficient and useful for sentiment analysis, spam detection, and topic classification. However, although Naïve Bayes is fast and easy to train, many comparative studies show that it often produces lower accuracy than SVM or Logistic Regression. For instance, in one study of public complaint classification, SVM obtained 0.84 accuracy, while Naïve Bayes obtained 0.77 accuracy [4]. In another comparative paper, Naïve Bayes scored lower than SVM and logistic Regression in performance [7]. This means Naïve Bayes is valuable when speed and simplicity are the primary priorities, but it may not be the top choice when higher predictive accuracy is needed.

Another traditional machine learning model is Random Forest which is frequently used in sentiment analysis studies. It combines multiple decision trees to get more stable classification and reduce overfitting. It is also helpful for handling complex data and many features. However, the results are mixed. One paper reported that Random Forest and SVM achieved the best overall results [8]. Another study reported Random Forest accuracy of 0.78, below SVM and XGBoost [4]. These findings tell that even though Random Forest can achieve better results; it may require careful tuning and may not be as consistently reliable as SVM for text sentiment tasks.

Feature extraction is an important task for sentiment analysis in the literature. Traditional machine learning models rely heavily on how text is expressed. TF-IDF is used more frequently because it shows important words while reducing the importance of most common words. Multiple studies use TF-IDF as the main representation for sentiment classification and show that it works especially well with SVM and logistic Regression. Some studies also examine n-grams, especially bi-grams, because they capture short word combinations that can improve sentiment detection [6]. For marketing companies, this is important because customer opinions often include short phrases such as “very good quality”, “not worth buying” or “highly recommend” which are better captured by n-grams than by single words alone.

Sentiment analysis is strongly related to marketing that companies keep track on social media posts, comments, product reviews for customer’s opinion to adjust their strategies, improve service quality. Online sentiment highly influences customer decision, so automated sentiment analysis is helpful for both businesses and customers. From marketing company’s point of view, sentiment analysis can benefit in customer satisfaction tracking, product feedback analysis and brand monitoring.

One single traditional model won’t be best for every dataset, but SVM performed consistently strong across many comparative studies, particularly when combined with

TF-IDF features. Depending on the dataset and feature extraction, Logistic Regression algorithm is also effective and in certain cases performs better. Random Forest is effective in handling complex data, but it is inconsistent for text analysis tasks and may overfit. Naïve Bayes is faster in training and easier to implement but generally less accuracy [4]. Evaluation metrics like accuracy, precision, recall and F1-score are used to evaluate performance of the algorithms. For marketing company, SVM is the best suitable traditional machine learning model for sentiment analysis, especially with TF-IDF feature extraction. SVM highly suitable model for because it performed consistently strongly in comparative studies, it achieves effectiveness for high-dimensional test data, it has balance between accuracy and computational efficiency. Logistic Regression can be considered as a second-best option if the company wants a simple and more interpretable alternative.

WORKFLOW AND METHODOLOGY:

4.1 Overview of Machine Learning Pipeline

Traditional Machine Learning workflow follows a structured pipeline, that is data collection, preprocessing, feature extraction, model development and evaluating the model. This is commonly used pipeline for sentiment analysis to transform raw textual data into meaningful predictions. Machine learning algorithms need numerical form of data, so before applying ML algorithm – should convert unstructured text into numbers. In this study, customer reviews are collected and pre-processed and then transformed into numeric data by using TF-IDF techniques. Machine learning model will be developed on that data and evaluation will be done by using standard metrics like accuracy, precision, recall and F1-score [2].

4.2 Data Description

The customer reviews dataset contains textual data like opinions about products or services labelled into sentiment categories (positive, negative or neutral). Review datasets are unstructured, so preprocessing is essential before analysis. Social media and e-commerce websites generate large volumes of the review data and making sentiment analysis essential for extracting insights. Such datasets are commonly used including movie reviews, product reviews and twitter data to evaluate sentiment classification models.

4.3 Data Preprocessing

Data preprocessing is an important step in sentiment analysis because reviews contain a lot of noise, so preprocessing can improve model performance and reliability.

Lowercasing – To prevent duplication of features and reduce dimensionality, all the text data is converted into lowercase which ensures consistency and reduces unnecessary complexity in feature space.

Removing Stopwords – Words like “is”, “and” and “and” which do not give much information to classification model. Focusing on meaningful words by removing stopwords can improve model efficiency.

Tokenization – Splitting sentence into individual words which is needed to convert sentences into processable units because model process the text at word level.

Cleaning Text – Deleting punctuations, numbers, special characters and irrelevant symbols which enhances the quality of input data to the model so model can perform better [9].

4.4 Feature Extraction

The input text must be transformed into numerical features after preprocessing. TF-IDF (Term Frequency – Inverse Document Frequency) – it gives priority to words based on the frequency according to the frequency and distinctiveness throughout the documents.

Term Frequency (TF) = Number of times a word appears in a document / Total number of words in the document

Inverse Document frequency (IDF) = $\log$ (Total number of documents / Number of documents containing the word)

TF-IDF = TF * IDF

A word with high TF-IDF means it occurs frequently in a document, and it found rarely in other documents.

4.5 Model Selection

For customer review sentiment analysis, Support vector machine is selected as the primary model because of its high performance in my literature review. In high-dimensional text data, SVM is a powerful supervised learning algorithm that performs effectively when paired with TF-IDF. Support Vector Machine achieves the highest accuracy compared to other classifiers in sentiment analysis [1]. SVM outperforms Naïve bayes in comparative analysis on review data [3]. SVM provides better results than Naïve bayes when applied to customer and airline reviews [10]. Furthermore, optimised SVM models achieve the best generalization performance after tuning [8]. SVM is chosen as the most appropriate model because of its high accuracy, robustness and suitability for the text classification.

Although deep learning models like CNN, LSTM can provide higher accuracy, traditional machine learning algorithms remain efficient, takes less computational resources, shorter training time for sentiment analysis tasks, particularly with limited resources [2].

4.6 Model Training

Provided customer review dataset is divided into training (80%) and testing (20%) sets. This split ensures that the model learns from one portion of the data while being fairly assessed on another. k-fold cross validation is used to the training set to enhance robustness. The training data is split into k subsets and then the model will be trained is trained on different folds. The mean result over all folds gives a more robust evaluation. The SVM model is trained to categorise the sentiment based on the extracted features by using techniques like TF-IDF.

4.7 Evaluation Metrics

To evaluate the performance of the model, accuracy, precision, recall and F1-score are the metrics commonly used in sentiment analysis. By using confusion matrix true positive (TP), true negatives (TN), false positives (FP) and false negatives (FN) are calculated which helps in providing deeper insights of model performance.

Accuracy – measures the overall correctness of predictions.
$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision – measures how many predicted positives are actually correct.
$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall – Measures how many actual positives are correctly identified.
$$\text{Recall} = \frac{TP}{TP+FN}$$

F1-Score – It is a combined score of correctness and completeness. Harmonic mean of precision and recall.

$$F1 = \frac{2 * (\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}$$

DISCUSSIONS AND RECOMMENDATIONS:

5.1 Discussion of Results

Based on the literature Support Vector Machine achieved better accuracy and better generalisation on unseen data, and the model gave better performance when evaluated with other models like Naïve Bayes, Logistic Regression in sentiment analysis. SVM handles high-dimensional data efficiently, which is common in sentiment analysis due to wide range of vocabularies because it is particularly effective for text classification and several studies confirm that SVM consistently achieves strong performance in sentiment analysis text classification tasks when compared to other traditional machine learning models [1]. Moreover, research tells that enhanced SVM models can perform very high accuracy, especially when paired with feature extraction techniques like TF-IDF [8]. This tells why SVM algorithm performed better in this study. Although Naïve Bayes is faster, it lacks accuracy due to its strong independence assumptions. Logistic Regression provides balanced results, but SVM outperforms it in handling complex decision boundaries in text data.

5.2 Justification of Chosen Model

In high dimensional text datasets SVM is known for its ability to find the optimal decision boundary which is hyperplane that improves classification accuracy. Although SVM is comparatively less effective than Naïve Bayes, it can handle large datasets effectively and performs efficiently with kernels and. When compared to non- SVM models, using a kernel reduces computational complexity also it practical for real-world applications like sentiment analysis for customer reviews of marketing company. It gives high accuracy in sentiment detection, reliable performance on large-scale data and ability to handle diverse customer opinions SVM achieves accurate and robust predictions, which are important for analysing text.

5.3 Practical Implementation

In analysing sentiment of customers review, the company can implement the SVM-based sentiment analysis system by collecting customer reviews from sources like websites or social media then Preprocess the text data like cleaning, tokenization, stopwords removal. After preprocessing convert text into numerical features using TF-IDF and then Apply the trained SVM model to classify sentiment. The system can be used for monitoring customer feedback in real time, can identify negative reviews quickly and improving products and services based on customer opinions. Studies highlight that sentiment analysis systems help organizations understand customer behaviour and improve decision-making processes [11].

5.4 Limitations

Although SVM algorithm performs well, it has some limitations like requires careful parameter tuning like regularization parameter, training time can increase with very large datasets, does not easily capture contextual meaning like sarcasm, dataset may not represent all real-world scenarios, presence of noisy text like different slangs and abbreviations affects accuracy and imbalanced data can bias predictions research indicates that sentiment analysis models often struggle with context understanding and informal language used in social media [11].

5.5 Recommendations

TF-IDF enhances feature representation by assigning importance to relevant words, which improves SVM performance in text classification tasks. Updating the model with new data regularly can analyse sentiment well and to adapt to changing customer language and trends. Using larger and more diverse datasets that Include data from multiple platforms can achieve highly accurate classification of sentiment.

CONCLUSION:

This study was aimed at applying traditional machine learning algorithms for sentiment analysis of customer review text data. Through the execution of preprocessing, feature extraction, and model comparison, the results showed that Support Vector Machine (SVM) paired with TF-IDF feature extraction achieves the optimal performance.

With in the evaluated models, SVM achieved greater predictive accuracy and robustness when dealing with high-dimensional text data. This is because that SVM is strongly effective in learning optimal decision boundaries, notably in sparse datasets generated through TF-IDF feature representation. Research also support that SVM demonstrates consistent performance sentiment classification tasks compared to other traditional models. Moreover, the use of TF-IDF improves the priority of meaningful words, enhancing the model's capability to classify between various sentiments.

As businesses progressively relying on sentiment analysis to interpret customer sentiment and enhance their products and services. The results of this study are important for real-world applications. marketing companies can enable automated analysis of large-scale feedback data and enable evidence-based decisions by

implementing an efficient and accurate model like SVM. For identifying insights from raw text data in multiple domains sentiment analysis has been widely acknowledged as an effective tool.

Finally, support vector machine is the perfect tradition machine learning model for sentiment analysis for customer reviews in the marketing company which is looking to minimise the computation cost and run their workflow on less resources.

REFERENCES:

- [1] G. Kaur and A. Sharma, "Comparison of Different Machine Learning Algorithms for Sentiment Analysis", *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, pp. 141–147. [Online].doi: 10.1109/ICSCDS53736.2022.9760846.
- [2] R. Chahar and A. K. Dubey, "Machine learning-based Sentiment Analysis: A Comprehensive Review", *2024 International Conference on Cognitive Robotics and Intelligent Systems (ICC - ROBINS)*, pp. 52–57. [Online].doi: 10.1109/ICC-ROBINS60238.2024.10533989.
- [3] S. Rana and A. Singh, "Comparative analysis of sentiment orientation using SVM and Naive Bayes techniques", *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, pp. 106–111. [Online].doi: 10.1109/NGCT.2016.7877399.
- [4] I. H. Hafidz, A. Sulistya and B. R. Lidiawaty, "Sentiment Analysis of Public Complaints: A Machine Learning Comparison of SVM, Naive Bayes, Random Forest, and XGBoost", *2025 International Conference on Advancement in Data Science, E-learning and Information System (ICADEIS)*, pp. 1–6. [Online].doi: 10.1109/ICADEIS65852.2025.10933382.
- [5] S. Mishra, et al., "Comparison of Machine Learning Techniques for Sentiment Analysis", *2023 3rd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS)*, pp. 184–191. [Online].doi: 10.1109/ACCESS57397.2023.10200806.
- [6] A. Sinha, et al., "Text Classification Using Machine Learning Techniques: Comparative Analysis", *2022 OITS International Conference on Information Technology (OCIT)*, pp. 102–107. [Online].doi: 10.1109/OCIT56763.2022.00029.
- [7] L. Maada, et al., "Sentiment Analysis: Machine Learning approaches comparison", *2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM)*, pp. 1–6. [Online].doi: 10.1109/WINCOM59760.2023.10322886.

- [8] R. Pandey, et al., "A Comparative Study of Random Forest, SVM, and Naive Bayes for Sentiment Analysis Optimization", *2024 IEEE 2nd International Conference on Innovations in High Speed Communication and Signal Processing (IHCSP)*, pp. 1–4. [Online].doi: 10.1109/IHCSP63227.2024.10959957.
- [9] A. Chauhan, A. Agarwal and R. Sulthana, "Performance Analysis of Machine Learning Algorithms and Feature Extraction Methods for Sentiment Analysis", *2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, pp. 1–8. [Online].doi: 10.1109/ICSES52305.2021.9633882.
- [10] A. M. Rahat, A. Kahir and A. K. M. Masum, "Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset", *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, pp. 266–270. [Online].doi: 10.1109/SMART46866.2019.9117512.
- [11] S. Sindhu, S. Kumar and A. Noliya, "A Review on Sentiment Analysis using Machine Learning", *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, pp. 138–142. [Online].doi: 10.1109/ICIDCA56705.2023.10099665.

FINAL GRADE

GENERAL COMMENTS

71 /100

Generic Assessment Criteria		Marks available
1. Engagement with Literature Skills		
.		
Good number and quality of sources in reference.		11
Referenced reasonably well throughout the document. There could be some improvement to the referencing through backing up points made in the text a little more with sources.		20
2. Knowledge and Understanding Skills		

Some good knowledge through the research shown in the text. 30

A little more backed up points through referencing would have been beneficial.

Reasonable critical analysis shown in the research part, showing both good and bad to each of the algorithms. 40

Bringing finetuning into the workflow would have been advantageous as this is an important stage to achieving the accuracy.

Overall some good knowledge and discussion for the ML models and process.

3. Cognitive and Intellectual Skills

Some good comparisons made between the algorithms. 25

A good workflow shown, however model finetuning would have been an advantageous addition here. 30

Reasonable recommendations based on researched evidence.

Good to see mechanics of the algorithms shown through formulae.

4. Practical Application Skills

0

5. Transferable Skills for Life and Professional Practice

A reasonably laid out report. Some more formatting and the addition of a cover page would have been beneficial. 10 5

A little more attention to the flow of text and sentences would have been advantageous.

A little repetition of process but overall good.

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9
