



Recommendation of a Sentiment Analysis ML model

Student Name: Mohammad Lakhi

Student ID: 2512430

Subject Code: ACCA7025

Subject Name: Application of Machine Learning

University of Wales Trinity Saint David

Date: 23rd October 2025

Abstract

This study examines the application of traditional methods of machine learning to sentiment analysis and suggests the best model for a marketing firm with small computational resources. In order to achieve a balance between accuracy, efficiency, and interpretability, it compares algorithms like Support Vector Machines (SVM), Naïve Bayes, and Logistic Regression. Key steps such as data pre-processing, feature extraction using TF-IDF, model training, and evaluation using accuracy, precision, recall, and F1-score are all covered in the straightforward workflow that is suggested. SVM is the most dependable and economical method for analyzing customer reviews, according to the results of recent research. The study comes to the conclusion that traditional models are still useful and efficient for small businesses, and that future research could concentrate on enhancing their adaptability through improved feature optimization and pre-processing.

TABLE OF CONTENTS

1	Introduction	1
2	Related Work in the Literature	3
3	Workflow methodology	5
4	Discussion and Recommendations	11
5	Conclusion and Future work	12
5.1	Conclusion	12
5.2	Future Scope	12
	REFERENCES	13

Chapter 1

Introduction

Sentiment analysis is a branch of natural language processing (NLP) which focuses on the identification and categorizing of emotions and opinions expressed in textual data. Its main function is to make the computer understand human sentiments from written text, from which it classifies whether the expression is positive, negative, or neutral [1][2].

As social media and e-commerce has been expanding and have become an important part of every human being, where they can express their feelings. So, this shows that the volume of public opinion and sentiments has increased significantly in the form of reviews, comments, and posts. These widely available data can be used by companies to improve customer satisfaction, and will also improve brand reputation and their financial growth [1][2].

Before the use of machine learning, sentiment analysis relied heavily on lexicon-based methods, in which predefined dictionaries of positive and negative words were used to identify emotions in text. These methods assigned opposite scores using manually crafted word lists and simple linguistic rules. While lexicon methods are useful for small and structured datasets, they frequently fail to capture contextual details like sarcasm, negation, and domain-specific language. As digital platforms expanded and customer feedback became more diverse and unstructured, rule-based systems struggled to remain accurate. This limitation created a need for machine learning approaches, which can automatically learn sentiment patterns from large amounts of text data and adapt more effectively to different contexts.

However, in today's time, deep learning approaches have demonstrated high performance in text classification tasks. But, they requires high computational resources and large datasets for training, which makes them unsuitable for small companies and

businesses [2]. In comparison to machine learning (ML) methods such as Navie Bayes (NB), Random Forest (RF), Logistic Regression, Gradient Boosting and Support Vector Machine (SVM) offer efficient alternatives for sentiment analysis while maintaining higher accuracy [5].

The key aim of this report is to provide an in-depth analysis of machine learning methods for a marketing company by evaluating and recommending a simple, cost-effective traditional algorithm for the sentiment analysis of customer reviews. To achieve this objective, a critical review of the literature on traditional machine learning algorithms will be conducted, to suggest a comprehensive and cost-efficient process for creating an accurate sentiment classification model. In the conclusion, to recommend a specific algorithm that effectively balances performance with minimal computational demands.

Chapter 2

Related Work in the Literature

By examining the existing literature is required to understand the performance, strengths, and drawbacks of traditional machine learning methods in sentiment analysis. This report focuses on research that compares various machine learning methods to suggest an appropriate and effective model for a marketing company with limited use of computational resources. This literature acknowledges that implementing an algorithm requires balancing between accuracy, computing cost, and interpret-ability [2].

Traditional machine learning methods are being used in sentiment analysis because of their efficiency and simplicity with average size of datasets. Styawati et al.[3] founded that Support Vector Machine (SVM) and Naïve Bayes classifiers are capable of performing sentiment analysis in social media textual data, indicating that preprocessed textual data can produce competitive outcomes without creating a complex model. Similarly, Anilkumar et al.[6] have used NB and SVM in product review data sets and found precision precision at minimal computing costs. These results shows that traditional machine learning methods are more effective and reliable for sentiment analysis, especially where computing cost efficiency is a priority.

Several research have reviewed traditional machine learning methods to find the best model for classifying customer and product review sentiments. In the Jannah et al.[4], SVM outperformed Naïve Bayes in terms of accuracy and consistency for marketplace reviews. Similarly, Henderi et al.[5] discovered that models such as SVM and logistic regression offer a better combination of performance and less computational cost. In order to predict customer sentiments on insurance product reviews, Hossain et al.[9] also used conventional machine learning techniques, such as Logistic Regression and Naïve Bayes, reporting competitive accuracy with short training times. Bhanu Prasad

et al.[8] came to a similar conclusion, Where SVM performs at high accuracy on Amazon product reviews. This finding demonstrate that SVM remains a solid and efficient approach for sentiment analysis in market review datasets.

While deep learning models like convolutional and neural networks have given higher accuracy in complicated sentiment analysis tasks, they require more processing time for larger datasets. Deep learning is less compatible for a small marketing companies with minimal infrastructure due to these limits. According to Monika et al.[11], traditional machine learning methods such as Naïve Bayes, Logistic Regression, and SVM provide faster training and uses lower resource while maintaining higher accuracy for sentiment analysis based on reviews. Similarly, Godia et al.[10] defines that traditional machine learning methods remain more interpretable and easier to deploy in production situations, which is vital for organizations that want immediate insights.

Despite the consistent performance of traditional machine learning methods, previous research has frequently focused solely on algorithm accuracy, with little emphasis on workflow optimization for small-scale deployment. According to Monika et al.[11], the pre-processing and feature extraction methods used have a significant impact on sentiment classification results. However, many existing studies are based on high-resource environments or ignore computational trade-offs. As a result, there is still a need for a straightforward and practical workflow that ensures accuracy and efficiency in real-world business situations. This report aims to address this gap by understanding findings from research and proposing a structured, cost efficient sentiment analysis workflow tailored to a marketing company with limited resources.

Chapter 3

Workflow methodology

This section describes the proposed machine learning process, which is used to create an accurate and efficient sentiment analysis model for classifying marketing companies review. The workflow is specifically designed to meet the client's needs for using traditional machine learning methods while having minimum computational resource usage. This methodology focuses on commonly used and proven techniques to ensure reliability and accuracy while keeping costs under control.

This proposed methodology will involve following key phases, which progress sequentially from initial data handling to final model assessment :

1. Data Collection and Preparation: Collecting and planning raw shopper reviews.
2. Text Pre-processing: involves cleaning and formatting the review text.
3. Feature extraction: converting text into a numerical format that is suitable for machine learning algorithms.
4. Data splitting: process of dividing a dataset to train and evaluate models.
5. Model Training and Hyperparameter Tuning: Creating and improving the selected traditional ML model.
6. Model evaluation: reviewing of final model's performance and efficiency.

This general pipeline structure is consistent with standard practices found in sentiment analysis research. Its an end-to-end sentiment analysis frameworks, emphasizing every key stages of the workflow [1][2][11].

1. Data Acquisition and Preparation:

The process begins by obtaining the shopper review data directly from the mar-

keting company, which serves as the primary input for the analysis. Once received, this data will be organized into a usable format, ensuring each review text is clearly identifiable. While the core analysis focuses on the review text itself, any accompanying basic information provided by the customer will be maintained if relevant. Before proceeding with detailed cleaning, a quick assessment of the data's volume and format will be performed to understand the scale of the task and identify any immediate structural issues that need addressing. This approach is consistent with previous research, in which data from online marketplaces or customer feedback systems is validated and structured prior to analysis. Studies such as Tan et al.[1], Jogi et al.[8], and Monika et al.[11] have demonstrated that systematic preparation improves data quality and reduces noise in subsequent modeling stages.

2. Text Pre-processing:

Collecting and organizing the raw review data, the text is prepared for analysis using a series of pre-processing operations. Text pre-processing is required to convert unstructured textual data into a consistent and readable format. This process usually involves converting all text to lowercase and removing punctuation, numerical characters, emojis, and stop words, which helps eliminate irrelevant information that could bias the model. This process is then used to break down sentences into individual words or tokens, followed by rooting to reduce words to their base forms. These steps ensure that the text input is uniform and semantically meaningful for future feature extraction and model training.

Effective pre-processing is critical for increasing model accuracy and reducing noise in sentiment classification. Previous research has shown that thoroughly cleaning and normalizing text data improves the performance of traditional machine learning classifiers. This stage is consistent with the practices recommended by Styawati et al.[3], Anilkumar et al.[6], and Monika et al.[11], who found that efficient pre-processing significantly improves model reliability while reducing computational overhead.

3. Feature Extraction:

Feature extraction is the conversion of textual data into numerical representations that capture the underlying meaning and significance of words or phrases. It helps machine learning algorithms recognize linguistic patterns by converting unstructured text into measurable features. This step is critical in sentiment analysis because it bridges the gap between human language and computational models, allowing algorithms to precisely analyse opinions and emotions expressed in text. Bag-of-Words and Term Frequency-Inverse Document Frequency are two popular feature extraction methods that quantify the importance of terms across documents and improve model learning and classification performance[1][2][6].

4. Data Splitting:

It is an important step in developing a reliable sentiment analysis model because it ensures that the trained algorithm can effectively generalize to new data. During this process, the dataset is divided into distinct subsets—typically training, validation, and testing sets—to objectively assess the model’s accuracy and performance. The training data is used to fit the model, the validation data helps with hyperparameter tuning, and the testing data evaluates the final model’s effectiveness. Depending on the size and complexity of the dataset, the common splitting ratios are 70:30 or 80:20 [1][2].

Maintaining class balance across these subsets is critical in sentiment analysis tasks to avoid bias, especially when dealing with imbalanced datasets with overwhelmingly positive or negative sentiments. Tan et al.[1] and Mao et al.[2] found that data partitioning is crucial for evaluating traditional Machine Learning classifiers like Naïve Bayes and Support Vector Machine under realistic conditions. Similarly, Anilkumar et al.[6] used an 80:20 train-test split for product review data, showing that careful data division contributes to stable accuracy and prevents overfitting.

5. Model Training and Hyperparameter Tuning:

The model training phase includes selecting a suitable traditional machine learn-

ing algorithm and applying it to the preprocessed and feature-extracted dataset. During this stage, the model learns to associate patterns in textual features with sentiment labels (positive, negative, or neutral). This report focuses on traditional machine learning algorithms like Naïve Bayes, Logistic Regression, and Support Vector Machine, which offer a balance of simplicity, interpret-ability, and efficiency [3][4][6].

Training involves inserting processed data into the chosen algorithm to reduce prediction error and improve model parameters. For example, Styawati et al.[3] and Jannah et al.[4] demonstrated that SVM and NB perform well in sentiment classification with small datasets, achieving high accuracy with minimal resource usage. Anilkumar et al.[6] showed that combining TF-IDF features with SVM or NB produces reliable classification results in product review datasets without requiring extensive computation.

Hyperparameter tuning is the process of adjusting key parameters to improve model performance. Changing the C parameter in SVM, smoothing parameter in Naïve Bayes, or regularization strength in Logistic Regression can greatly impact accuracy and generalization. Henderi et al.[5] point out the importance of this process in preventing overfitting and achieving optimal performance within limited computational budgets. The model's parameters can be systematically optimized using techniques such as grid search or cross-validation to produce consistent and interpretable results appropriate for small-scale business applications.

6. Model evaluation:

It is crucial to evaluate a model's performance on unseen data after it has been trained and optimized. Instead of merely memorizing the training examples, this step guarantees that the model can generalize effectively. A set of common performance metrics, including accuracy, precision, recall, and the F1-score, are commonly used in sentiment analysis evaluation [2].

- **Accuracy:** The percentage of reviews that are correctly classified across all samples in the dataset is known as accuracy. It offers a broad picture of

the model's performance, though it might not accurately represent results in cases where class distributions are not uniform. For example, accuracy by itself may overestimate performance in customer reviews where positive sentiments predominate. Consequently, it needs to be enhanced by additional measures.

- **Precision:** Precision measures the number of reviews that were actually positive or negative as predicted. When companies depend on results to make decisions about marketing or services, a high precision score indicates that the model produces fewer false positive predictions.
- **Recall:** The model's recall measures its ability to recognize every review that falls into a specific sentiment class. The majority of true positive samples are successfully captured by a model with a high recall. This guarantees that all customer concerns are successfully identified in the context of customer reviews.
- **F1-score:** Precision and recall are balanced into a single performance value by the F1-score, which is the combined average of those two metrics. When there are uneven sentiment classes in the dataset, it is especially helpful. A model with a high F1-score balances accuracy and generalization, performing consistently well across various sentiment types.
- **Confusion matrix:** The model's predictions are visually summarized by a confusion matrix, which displays the proportion of correctly classified versus incorrectly classified reviews. It helps in further classifier improvement and is helpful in detecting patterns in model errors, such as whether neutral reviews are frequently confused for positive ones.

Workflow summary : This project's overall workflow follows to a well-organized and effective cycle that aims to guarantee accurate sentiment classification at a low computational cost. The first step in the process is data preparation and collection, which involves collecting and arranging customer reviews. After that, the raw text is put through pre-processing, which includes data quality-enhancing cleaning procedures

like stop-word removal, punctuation filtering, and normalization.

After pre-processing, textual data is transformed into numerical form using feature extraction techniques like TF-IDF or Bag-of-Words, which allows machine learning models to process the data efficiently. To ensure balanced model evaluation and prediction, the dataset is subsequently divided into training and testing subsets. Algorithms like Naïve Bayes, Support Vector Machine, and Logistic Regression are used to extract sentiment patterns from the features during model training. Simple optimization methods like grid search are used to adjust hyperparameter in order to improve performance without using a lot of processing power.

Finally, the model's efficiency is evaluated using confusion matrix, F1-score, recall, accuracy, and precision. This structured approach guarantees that the model will continue to be lightweight, interpretable, and suitable for use by small businesses with limited technical understanding [6].

Justification: According to the reviewed literature, Support Vector Machine (SVM) is the best algorithm for this project. SVM consistently achieves high accuracy in text classification and effectively manages high-dimensional feature spaces, making it ideal for sentiment analysis tasks.[4][5].

Jannah et al.[4]and Bhanu et al.[8] found that SVM outperformed Naïve Bayes and other traditional methods for product review data. Its robustness and ability to generalize across different data distributions make it suitable for marketing companies looking to analyze customer feedback without spending high infrastructure costs.

This chapter explained the entire workflow to build a sentiment analysis model with traditional machine learning techniques. Each phase, from data preparation to model evaluation, was designed to ensure accuracy and reliability while minimising computational demands. The chosen algorithm, SVM, provides a good balance between performance and simplicity, making it an appropriate choice for small or emerging marketing firms.

Chapter 4

Discussion and Recommendations

The findings from the reviewed literature and the proposed workflow clearly points that traditional machine learning methods are still highly effective for sentiment analysis, especially in environments with limited computational resources. Traditional machine learning models are easier to train, interpretable, and deploy than deep learning approaches, which require large datasets and powerful hardware. They also maintain competitive accuracy levels.

Among the methods tested, the Support Vector Machine (SVM) consistently demonstrated the best balance of accuracy, interpret-ability, and efficiency. Jannah et al.[4], Anilkumar et al.[6], and Bhanu et al.[8] found that SVM outperforms Naïve Bayes and Logistic Regression in classifying product and customer review sentiments. Its ability to handle high-dimensional data with a relatively small dataset makes it ideal for small marketing firms seeking practical sentiment analysis solutions.

As a result, Its recommends using a traditional ML workflow based on SVM with TF-IDF feature extraction and standard pre-processing steps. This workflow ensures consistent results at a low computational cost, aligning with the company’s operational capacity. Evaluating performance using metrics like accuracy, precision, recall, and F1-score will help to validate the model’s consistency over time.

Chapter 5

Conclusion and Future work

5.1 Conclusion

This report concludes that traditional machine learning methods, particularly Support Vector Machine (SVM) combined with TF-IDF feature extraction, are the most effective and resource-efficient approach to sentiment analysis of shopper reviews. The proposed workflow enables accurate, interpretable, and low-cost implementation suitable for small marketing firms. By implementing this model, businesses can gain meaningful insights from customer feedback without the need for advanced infrastructure, ultimately improving their decision-making and customer engagement strategies.

5.2 Future Scope

Building on the proposed workflow, future research may focus on improving the performance and adaptability of traditional machine learning models such as SVM via refined pre-processing and feature optimization. Using combination methods with classifiers like Logistic Regression, Naïve Bayes, and SVM could improve sentiment classification consistency across different product domains. Expanding the dataset to include a variety of review sources, such as social media and multilingual feedback, would improve the model's robustness. Furthermore, incorporating automated hyperparameter tuning and cross-validation techniques could ensure long-term accuracy and efficiency while minimizing computational requirements.

References

- [1] K. L. Tan, C. P. Lee, and K. M. Lim, “A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research,” *Applied Sciences*, vol. 13, no. 7, p. 4550, Apr. 2023, doi: <https://doi.org/10.3390/app13074550>.
- [2] Y. Mao, Q. Liu, and Y. Zhang, “Sentiment analysis methods, applications, and challenges: A systematic literature review,” *Journal of King Saud University. Computer and information sciences*, vol. 36, no. 4, pp. 102048–102048, Apr. 2024, doi: <https://doi.org/10.1016/j.jksuci.2024.102048>.
- [3] Styawati, A. R. Isnain, N. Hendrastuty, and L. Andraini, “Comparison of Support Vector Machine and Naïve Bayes on Twitter Data Sentiment Analysis,” *Jurnal informatika/Jurnal Informatika: Jurnal Pengembangan IT*, vol. 6, no. 1, pp. 56–60, Jan. 2021, doi: <https://doi.org/10.30591/jpit.v6i1.3245>.
- [4] Z. B. Jannah and K. Kusnawi, “Comparison of Naïve Bayes and SVM in Sentiment Analysis of Product Reviews on Marketplaces,” *Sinkron*, vol. 8, no. 2, pp. 727–733, Mar. 2024, doi: <https://doi.org/10.33395/sinkron.v8i2.13559>.
- [5] H. Henderi and Q. Siddique, “Comparative Analysis of Sentiment Classification Techniques on Flipkart Product Reviews: A Study Using Logistic Regression, SVC, Random Forest, and Gradient Boosting,” *Journal of Digital Market and Digital Currency*, vol. 1, no. 1, pp. 21–42, Jun. 2024, doi: <https://doi.org/10.47738/jdmdc.v1i1.4>.
- [6] C. Anilkumar, S. Kanchana, and S. B. Kumar, “Sentimental Analysis on Product Reviews Using Support Vector Machine and Nave Bayes,” *Applied and Computational Engineering*, vol. 2, no. 1, pp. 1067–1073, Mar. 2023, doi: <https://doi.org/10.54254/2755-2721/2/20220586>.
- [7] P. G. Bhanu 1, K. M. V 1, O. Chandana, Anjali 1, and S. and, “SENTIMENT ANALYSIS OF CUSTOMER PRODUCT REVIEWS USING MACHINE LEARNING,” *Turkish Journal of Computer and Mathematics Education*, vol. 14, no. 3, pp. 178–188, Available: <https://www.proquest.com/scholarly-journals/sentiment-analysis-customer-product-reviews-using/docview/2837151549/se-2>
- [8] R. K. Jogi et al, ”SUPERVISED MACHINE LEARNING APPROACH FOR SENTIMENT ANALYSIS ON AMAZON PRODUCT REVIEWS,” *International Journal of Communication Networks and Information Security*, vol. 17, (3), pp. 226–230, 2025. Available: <https://www.proquest.com/scholarly-journals/supervised-machine-learning-approach-sentiment/docview/3232790520/se-2>.
- [9] M. S. Hossain and M. F. Rahman, “Customer Sentiment Analysis and Prediction of Insurance Products’ Reviews Using Machine Learning Approaches,” *FIIB Business Review*, vol. 12, no. 4, p. 231971452211157, Aug. 2022, doi: <https://doi.org/10.1177/23197145221115793>.

- [10] Adarsh Godia and L. K. Tiwari, “Sentiment Analysis and Classification of Product Reviews: A Comprehensive Study Using NLP and Machine Learning Techniques,” 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), pp. 1247–1252, Mar. 2024, doi: <https://doi.org/10.1109/icaccs60874.2024.10717296>.
- [11] P. Monika, C. Kulkarni, N. Harish Kumar, S. Shruthi, and V. Vani, “Machine learning approaches for sentiment analysis,” International journal of health sciences, pp. 1286–1300, Apr. 2022, doi: <https://doi.org/10.53730/ijhs.v6ns4.6119>.