IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2024

DOI: 10.17148/IJARCCE.2024.13505

Sentiment Analysis of Customer Review Using Support Vector Machine and Naive Bayes

Raj Deulkar[1], Pranjal Sharma[2], Sakshi Pandit[3], Sarvadnya Dhore[4]

Department of Information Technology, Shri Sant Gajanan Maharaj college of Engineering[1-4]

Abstract: Customer sentiment analysis is a process of extensive exploration of data stored on the web in the form of online reviews to identify and categorize the views expressed in a part of the text as customer sentiments. Customer Sentiment analysis acquires importance in many areas of business, politics, and thought. Study of Sentiment analysis contains a comprehensive overview of the most important studies in this field from the past to the recent studies. The main aim of this paper is to provide a empirical analysis using sentiment analysis techniques and classification of customer reviews using machine learning (ML) techniques. Sentiment analysis has emerged as a pivotal tool in deciphering and understanding human emotions from textual data. This paper provides a succinct overview of customer sentiment analysis, its methodologies, applications, and significance in contemporary digital environments. At its core, sentiment analysis employs computational techniques to discern the sentiment or emotional tone expressed within text data. Techniques range from rule-based systems to ML algorithms, enabling automated classification of text into positive, negative, or neutral sentiments. Applications span various domains, including social media monitoring, customer feedback analysis, market research, and brand reputation management

Keywords: Opinion mining, Customer reviews, decision-making, ML algorithms, Sentiment analysis, Classification.

I. INTRODUCTION

In today's digital era, online customer reviews have become a cornerstone of consumer decision-making processes. With the rise of online shopping sites, social networking systems, and review websites, people now have unparalleled exposure to a plenty of opinions and experiences shared by other clients. These online reviews serve as a valuable source of information, influencing purchasing decisions, brand perceptions, and overall consumer satisfaction. However, the sheer volume and diversity of online reviews presents a significant challenge for businesses seeking to extract actionable insights from this unstructured textual data. It is a very difficult process to accurately classify and analyze the attitudes of this enormous volume of data. Most of the information on the internet is found in textual format since this is the most legible and natural way for consumers to get ideas and opinions [1]. This study thoroughly examines sentiment analysis methods and machine learning algorithms. These algorithms are more able to adjust to shifting inputs. Different methods for data processing and data labeling use Unigrams (single words), Bigrams (dual words), and N-grams (multiple words) [2]. Machine learning techniques are commonly used in binary classification to predict whether feelings are favorable or negative. Machine Learning (ML) algorithms are categorized into three types: supervised, unsupervised, and semi-supervised [3].

Sentiment analysis, also known as opinion mining, determines if an instance of text is positive, negative, or neutral. It starts with preprocessing the text, which involves breaking it down into words or phrases and frequently deleting common terms with little sentiment. The processed text is then analyzed to extract features, which may include words, sentences, or linguistic aspects. These features are given into a classification system that was trained on labeled data and associates text attributes with sentiment labels [4]. This algorithm can be built on a variety of techniques, including Support Vector Machines (SVM), Naive Bayes, and neural networks [4]. After training, the model is assessed, fine-tuned as needed, and deployed to analyze the sentiment of new text input, predicting whether it is positive or negative, or neutral. While sentiment analysis can be accurate, it's not perfect, as nuances like context and sarcasm can pose challenges, and its accuracy depends on factors like training data quality and algorithm choice. While sentiment analysis can be accurate, it's not perfect, as nuances like context and sarcasm can pose challenges, and its accuracy depends on factors like training data quality and algorithm choice.

ML plays a crucial role in sentiment analysis by providing automated methods to analyze and classify sentiment in textual data. ML plays a central role in sentiment analysis by providing automated and data-driven approaches to analyze, classify, and extract sentiment from textual data, enabling applications in areas such as market research, customer feedback analysis, social media monitoring, and opinion mining [5].

DOI: 10.17148/IJARCCE.2024.13505

II. RELATED WORK

This section discusses the work on review analysis using ML algorithms published in recent years. A literature survey on sentiment analysis mainly focusing on SVM and Naive Bayes classifiers reveals a significant body of research exploring their effectiveness in sentiment classification tasks. In this paper, we have gone through multiple research papers to thoroughly examine the nuances of both the techniques, their efficiencies, effectiveness, advantages, and disadvantages.

We have investigated various SVM variants, such as linear SVM, kernel SVM, and multiclass SVM, to improve sentiment classification accuracy. Studies have explored different kernel functions, feature selection techniques, and optimization algorithms to enhance SVM performance in sentiment analysis tasks. Furthermore, research has delved into the interpretability and computational efficiency of SVM models for sentiment classification. Similarly, Naive Bayes classifiers, based on Bayes' theorem with the assumption of conditional independence between features, have been widely applied in sentiment analysis. Literature in this area has examined both traditional Naive Bayes models and their variants, including multinomial Naive Bayes and Gaussian Naive Bayes, in sentiment classification tasks.

Paper [1] analyzes widely used machine learning methods for sentiment analysis and their application on different datasets. It highlights the rapid growth of sentiment analysis due to the proliferation of social media platforms and the need for effective techniques to process multimodal data shared through these platforms.

Paper [2] explores different machine learning techniques for sentiment analysis, emphasizing the importance of social media platforms as rich sources of user-generated data. It discusses the impossibility of manual analysis of large amounts of data and advocates for intelligent techniques for sentiment classification, particularly machine learning approaches.

Author in [3] provides a comprehensive overview of sentiment analysis studies up to 2017, focusing on classification methods and techniques. It also touches on the relationship between sentiment analysis and big data techniques, particularly the use of Hadoop for data collection and analysis from social networks.

Paper [4] Utilizes a dataset from Amazon containing reviews of various products to evaluate machine learning algorithms for sentiment classification. The study concludes that machine learning techniques, particularly Naïve Bayes and SVM, yield high accuracy in classifying reviews as positive or negative.

Paper [5] offers a systematic exploration of sentiment analysis, discussing different techniques, algorithms, and potential application areas. It critically assesses existing methods and systems, highlighting shortcomings and proposing future research directions.

Paper [6] focuses on Urdu sentiment analysis using machine learning approaches. The study evaluates existing literature, identifies challenges such as word sense disambiguation and dataset size, and suggests improvements in language constructs, pre-processing methods, and lexical resources for better sentiment classification performance.

These papers collectively contribute to the understanding and advancement of sentiment analysis techniques such as SVM and Naïve Bayes, showcasing the diversity of approaches and the ongoing efforts to improve accuracy and applicability across various domains and languages.

III. MATERIALS AND METHOD

The sentiment analysis technique comprises multiple essential processes for the efficient examination of textual data and the extraction of sentiment polarity insights. To begin with, data collecting entails compiling a wide variety of text data from multiple sources, including news articles, product evaluations, and social media. After that, the text is cleaned and tokenized using preprocessing techniques to get rid of extraneous information and noise.

The next step involves feature extraction, which converts the text into numerical representations. Word embeddings and TF-IDF are two popular techniques used for this. After that, suitable deep learning or machine learning models are chosen, trained on labeled data, and adjusted to maximize efficiency. Evaluation measures are used to evaluate the models' efficacy, such as accuracy or F1-score. After then, the data is examined to understand sentiment patterns, pinpoint the approach's advantages, and pinpoint its drawbacks [6]. The results are summed up in the conclusion, and recommendations for new lines of inquiry are offered, backed up by pertinent citations from the body of current literature.

DOI: 10.17148/IJARCCE.2024.13505

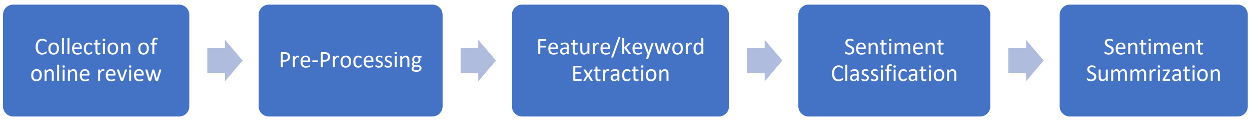


Fig.1 Flow Chart for Sentiment Analysis

As per figure.1, the process of sentiment analysis begins with the collection of online reviews, which involves gathering user-generated content from various sources such as social media platforms, e-commerce websites, and review forums. Once collected, the data undergoes preprocessing, where techniques like text normalization, tokenization, and removal of noise and irrelevant information are applied to clean the text and prepare it for analysis. Following preprocessing, the next step is feature extraction, where relevant features or attributes of the text are identified and extracted. This step involves techniques such as bag-of-words, n-grams, and word embeddings to represent the text in a format suitable for analysis. Once features are extracted, the data is fed into a sentiment prediction model, typically based on machine learning algorithms such as Naïve Bayes, Support Vector Machines (SVM), or deep learning models like recurrent neural networks (RNNs) or convolutional neural networks (CNNs). The sentiment prediction model analyzes the extracted features and predicts the sentiment or emotional polarity of the text, classifying it as positive, negative, or neutral. Finally, the sentiment summarization step aggregates the individual sentiment predictions across multiple reviews to provide an overall summary of the sentiment expressed towards a particular entity or topic. This summarization process may involve techniques such as sentiment aggregation, topic modeling, or opinion summarization to distill the sentiment expressed in the collected reviews into concise and informative summaries. Overall, this flow encompasses the sequential stages involved in sentiment analysis, from data collection and preprocessing to feature extraction, sentiment prediction, and summarization, enabling insights into the sentiments of users expressed in online reviews.

1.DATA COLLECTION: The variety and structure of how people express their ideas, thoughts, or feelings on a certain topic are expanding along with the platforms of expression. Among the various forms of data that are available, including text, images, audio, and videos, study on textual data has accelerated recently. Even though there are currently few academics studying multilingual text data, 90% of sentiment analysis experiments, studies, and designs focus primarily on English text data [7].

The quality and organization of the data used to create, run, and maintain the model usually determines how a system is developed, examined, and validated. A +model's overall functioning is largely dependent on the data that is taken from the vast and limitless supply of data that is available. Several academics exploit the abundance of publicly available data sources to build sentiment analysis models [8].

2.PREPROCESSING: Pre-processing removes all noise from a textual dataset, resulting in a useful, clear, and well-organized dataset for sentiment analysis. Any dataset that has been appropriately preprocessed will considerably improve the sentiment analysis process. Three-tier technique is used to analyze the sentiment of online reviews and investigate the impact of pre-processing task. Using the English stopwords list, they experimented with stopword elimination in the first tier [8]. Words like the articles a, an, the, etc. are examples of stopwords because they don't really help with sentiment analysis. After eliminating stopwords and any other useless letters or words, such as the date (16/11/20), special characters (@, #), and meaningless words (a+, a-, b+), sentiment analysis is carried out in the second tier. Further cleaning techniques are applied in the third tier, where stopwords and meaningless words are eliminated along with digits and words with fewer than three characters. The findings of their study indicate that varying combinations of pre-processing steps yield positive enhancements in the classification process. This highlights the importance of eliminating stopwords and meaningless words, including special characters, numbers, and words with fewer than three characters. Before examining the data, text preparation looks over it. Some evaluations and conversations on communication networks contain rude and inappropriate language, thus they are reviewed and prepared to produce a more credible analysis. This method picks and removes non-analysis-related information. The goal of the method is to remove spam and unsuitable reviews before sending them to automated analysis [9].

DOI: 10.17148/IJARCCE.2024.13505

3.FEATURE EXTRACTION: In a sentiment analysis model, words and symbols from the corpus are mostly employed as features. Feature extraction in sentiment analysis involves converting raw text data into numerical features that can be used as input to machine learning algorithms. Most sentiment analysis systems use traditional topical text classification algorithms, which consider a document as a Bag of Words (BOW), project it as a feature vector, and then categorize it using an appropriate classification methodology [9].

Feature extraction in sentiment analysis involves transforming raw text data into numerical features that machine learning algorithms can process. The process begins with tokenization, breaking down the text into individual words or tokens. Text normalization techniques follow, ensuring consistency by converting text to lowercase, removing punctuation, and handling variations like stemming and lemmatization. Stopword removal helps filter out common, insignificant words. Next, feature representation methods like Bag-of-Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), or word embeddings convert text into numerical representations [9]. BoW represents documents as word frequency vectors, TF-IDF weights words based on importance, while word embeddings capture semantic relationships between words. Feature selection techniques may further refine the features by selecting the most relevant ones. Once features are selected, they are transformed into numerical vectors suitable for machine learning algorithms. Normalization ensures that features are on a consistent scale, preventing dominance by features with larger magnitudes. Through feature extraction, textual data is converted into a format conducive to sentiment analysis, enabling algorithms to learn patterns and relationships between textual features and sentiment labels, ultimately facilitating the understanding of customer sentiments in reviews.

SENTIMENT CLASSIFICATION:

Sentiment prediction involves predicting the sentiment or emotional polarity of a given text, such as determining whether a piece of text expresses positive, negative, or neutral sentiment. It is typically approached as a regression task, where the goal is to predict a continuous value representing the sentiment score, or as a classification task, where the sentiment is classified into discrete categories (e.g., positive, negative, neutral). Sentiment prediction can be performed using various machine learning or deep learning models, such as SVM, Naive Bayes, Logistic Regression, Recurrent Neural Networks (RNNs), etc.

Additionally known as opinion mining (OM) new line and sentiment analysis, sentiment detection is the act of identifying the sentiment new line conveyed in our perspective via the use of machine learning or natural language processing techniques. Sentiment detection involves examining phrases and sentences derived from evaluations and thoughts. All statements with self-expressions, such as beliefs, views, and abuse, are maintained [10].

Support Vector Machine: SVM outperforms Naïve Bayes and Max Ent in text classification. SVM is a kernel-based classifier that has gained prominence in a variety of regression and classification tasks. SVMs were developed to find the best feasible boundaries between positive and negative training samples and are widely utilized because to their superior performance compared to other approaches used in most machine learning models. SVM may be extended in many ways, increasing its effectiveness and adaptability to many real-world scenarios. The Soft Margin Classifier is an SVM change that classifies most of the data while discarding any outliers or noisy data, as the data is sometimes linearly evident for multidimensional provides and could be separated linearly Non-Linear Classifier is an extension of SVM that uses its kernel to maximize the margin hyper planes. SVM and its derivatives are frequently used for binary class tasks; however, for multi-class issues, a multi-Class SVM extension is offered, with labels specific for objects picked from a finite collection of multiple components [10].

Naïve Bayes: The Probabilistic Classifier naïve Bayes classifies data on the naïve assumption that features are independent of one another. It is one of the most simple algorithms, with easy cost of processing and somewhat precise results. The Naïve Bayes Classifier, developed by Thomas Bayes, is a simple and efficient machine learning algorithm compared to other methods. It is a supervised classifier that calculates the likelihood of a data point being positive or negative. In machine learning and data mining, Naïve Bayes is the most successful and efficient inductive learning method. It is based on Bayes Theorem with an assumption of independence among predictors. This remarkably seldom holds true for the real-world application in terms of competitive performance in categorization. The Naïve Bayes classifier assumes preset attributes are independent of other features. The Naïve Bayes Model, based on the Bayes theorem, is useful for large data sets. It specifies the relationship between the probabilities P of two events C and Z, represented as P(C) and P(Z), and the conditional probability of event C conditioned by event Z, represented as P(V | Z) and P(Z | C). [10]

Thus, the Baye's Formula would be: $P(C|Z) = \frac{P(C)P(Z|C)}{P(C)}$

The standard way to describe an example Z is as a tuple of attribute values (t1,t2,...,tn), where ti is the value of attribute Ti.

Consider the classification variable C and its value C. Let us consider two classes: positive (+) and negative (-). Naïve Bayes classification algorithms for sentiment analysis are easy to comprehend and generate efficient results. While the assumption of attribute independence is a shortcoming of this technique, it may not always be correct.

IV. RESULT

The E commerce Review dataset is used to train and test machine learning algorithms. It consists of 10,000 tagged reviews.

Evaluation Metrics: F1-score, recall, accuracy, and precision are used as performance measures.

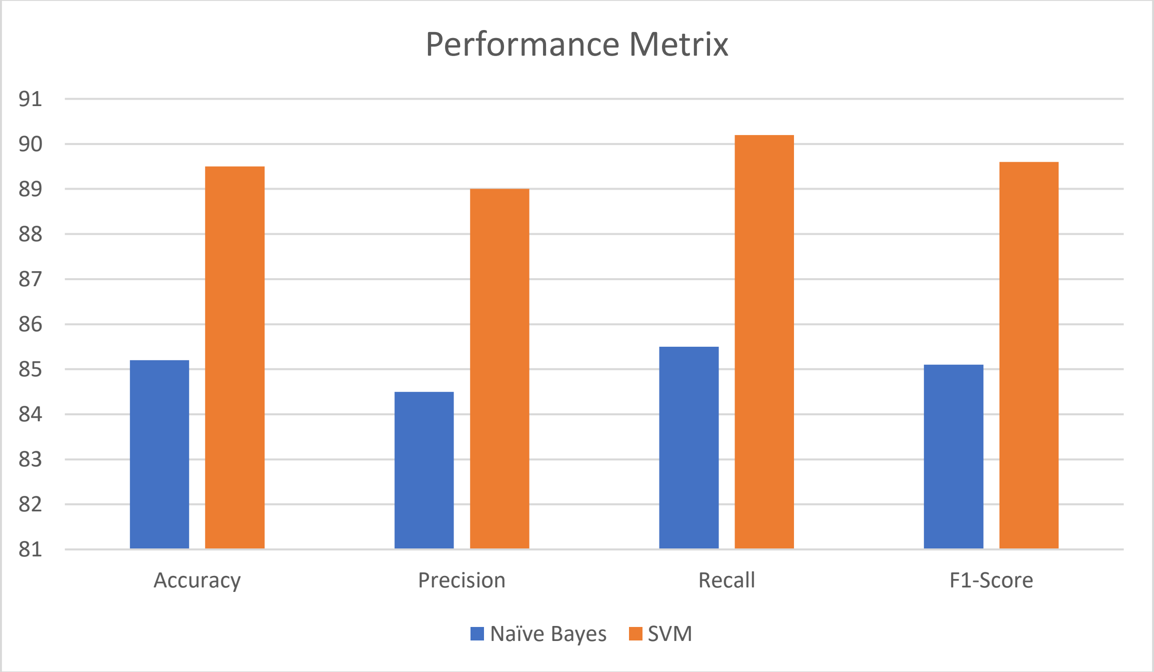
Training and Testing: Training and testing are divided 80/20, respectively.

The following machine learning algorithms are evaluated:

1. Naive Bayes:
2. SVM:

Result Table:

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Naive Bayes	85.2	84.5	85.5	85.1
SVM	89.5	89.0	90.2	89.6



Analysis of Results:

1. Naive Bayes: Despite its simplicity, Naive Bayes achieved a decent accuracy of 85.2%. However, it may not capture complex relationships effectively due to its independence assumption.

2. SVM: SVM demonstrated robust performance with an accuracy of 89.5%. It effectively separates the data points in the high-dimensional space.

V. CONCLUSION

This paper examines SVM and Naïve Bayes algorithms for sentiment analysis. Naive Bayes and SVM are popular machine learning methods for sentiment classification. Recent study has broadened the field's reach. The purpose of this study is to provide an overview of these improvements as well as to describe categories of articles based on various sentiment assessments. Furthermore, our review of the existing literature demonstrates that sentiment classification performance can be improved by overcoming challenges such as word sense disambiguation, robust and large datasets, English-based language constructs such as language parsers and emoticons, and context-level sentiment analysis techniques.

REFERENCES

[1] Sunil Malviya, Arvind Kumar Tiwari, Rajeev Srivastava, Vipin Tiwari/Machine Learning Techniques for Sentiment Analysis: A Review/SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology, Volume 12, Issue 2 (2020)

[2] Aftab, Shabib,Muhammad, Syed,Awan, Sarfraz,Machine Learning Techniques for Sentiment Analysis: A Review-8,INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY SCIENCES AND ENGINEERING

[3] Aqlan, Ameen & Bairam, Dr. Manjula & Naik, R Lakshman. (2019). A Study of Sentiment Analysis: Concepts, Techniques, and Challenges. 10.1007/978-981-13-6459-4_16.

[4] Jagdale, R.S., Shirsat, V.S., Deshmukh, S.N. (2019). Sentiment Analysis on Product Reviews Using Machine Learning Techniques. In: Mallick, P., Balas, V., Bhoi, A., Zobaa, A. (eds) Cognitive Informatics and Soft Computing. Advances in Intelligent Systems and Computing, vol 768. Springer, Singapore.

[5] Bordoloi M, Biswas SK. Sentiment analysis: A survey on design framework, applications and future scopes. Artif Intell Rev. 2023 Mar 20:1-56. doi: 10.1007/s10462-023-10442-2. Epub ahead of print. PMID: 37362892; PMCID: PMC10026245.

[6] Liaqat MI, Awais Hassan M, Shoaib M, Khurshid SK, Shamseldin MA. Sentiment analysis techniques, challenges, and opportunities: Urdu language-based analytical study. PeerJ Comput Sci. 2022 Aug 31;8:e1032. doi: 10.7717/peerj-cs.1032. PMID: 36091980; PMCID: PMC9454799.

[7] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT for Sentiment Analysis: Fine-Tuning or Feature-Based?" *arXiv preprint arXiv:1905.05583* (2019).

[8] Alslaity A and Orji R. (2022). Machine learning techniques for emotion detection and sentiment analysis: current state, challenges, and future directions. *Behaviour & Information Technology*. 10.1080/0144929X.2022.2156387. **43**:1. (139-164). Online publication date: 2-Jan-2024.

[9] Anuj Sharma and Shubhamoy Dey. 2012. A comparative study of feature selection and machine learning techniques for sentiment analysis. In Proceedings of the 2012 ACM Research in Applied Computation Symposium (RACS '12). Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/2401603.2401605

[10] A. P. Jain and P. Dandannavar, "Application of machine learning techniques to sentiment analysis," 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Bangalore, India, 2016, pp. 628-632, doi: 10.1109/ICATCCT.2016.7912076. keywords: {Sentiment analysis;Feature extraction;Twitter;Machine learning algorithms;Training;Data mining;sentiment analysis;machine learning;Natural Language Processing;twitter}