SENTIMENT ANALYSIS USING DIFFERENT MACHINE LEARNING MODELS: A STUDY FOR THE PREDICTION OF CUSTOMER'S REVIEW

Kallal Banerjee, Swami Vivekananda University
Sweety Sarkar, Swami Vivekananda University

ABSTRACT

Nowadays, the world is becoming digitalized. e-commerce is ascending in this digitalized world through the availability of products within reach of customers. Furthermore, e-commerce websites allow people to convey their thoughts and feelings. People are increasingly relying on the experiences of other customers. Our opinions and purchasing decision-making are affected by the experience of others and their feedback about products. We always ask others about their opinion to get benefit from their experience; hence, the importance of reviews has grown. However, it is almost impossible for customers to read all such reviews; therefore, sentiment analysis is essential in analyzing them. This study proposes a sentiment analysis to predict the polarity of Amazon baby product dataset reviews using supervised machine learning algorithms. Further, it will allow companies to improve their products by knowing customers' opinions and needs. Amazon is one of the e-commerce giants that people use daily for online purchases where they can read thousands of reviews dropped by other customers about their desired products. These reviews provide valuable opinions about a product such as its property, quality, and recommendations, which helps the purchasers understand almost every detail. This project considers the sentiment classification problem for online reviews using supervised approaches to determine the overall semantics of customer reviews by classifying them into positive and negative sentiments.

Keywords: K-Nearest Neighbor, Logistic Regression, Support Vector Machine, Supervised Machine Learning, Un-supervised Machine Learning.

INTRODUCTION

Sentiment analysis, a machine learning technique in natural language processing (NLP), is vital in extracting sentiments from textual data sources such as social media, surveys, and e-commerce reviews. It enables entrepreneurs to gain insights into customer perspectives and understand product satisfaction factors. In recent years, sentiment analysis has evolved to encompass various classification approaches, including polarity-based classification, intent or emotion detection, and aspect-level sentiment analysis (Ingle et al. 2015). Several studies have explored sentiment classification strategies, particularly in the context of social media platforms like Twitter (Sahayak et al., 2015). These strategies involve machine learning algorithms and models to analyze and classify sentiments accurately. Additionally, research has focused on text mining and consumer feedback classification, leveraging techniques such as Recurrent Neural Networks (RNNs) with Gated Recurrent Units (GRUs) and paragraph vector integration. To contribute to this field, our project aims to develop a robust sentiment analysis system capable of accurately predicting sentiments in text data. We will curate a diverse dataset comprising labeled examples, preprocess the data, and employ various feature engineering techniques, including word embeddings, to capture

Citation Information: Banerjee K., Sarkar S.,(2024). Sentiment analysis using different machine learning models: A study for the prediction of customer's review. *Journal of Legal, Ethical and Regulatory Issues, 27*(1), 1-10.

semantic information effectively. Our machine-learning model will be trained and evaluated using appropriate metrics to ensure its accuracy and generalizability (Agarwal et al., 2012).

By deploying our sentiment analysis system, we seek to provide a practical tool that empowers businesses, researchers, and individuals to analyze real-time data from different online media sources (Chandrakala, 2012). The system will facilitate understanding customer opinions, brand reputation management, and market research (Cambria et al., 2017). We aim to contribute to advancing sentiment analysis techniques and enable informed decision-making based on sentiment analysis of textual data.

Motivation

This study aims to comprehensively describe, demonstrate, and assess supervised machine-learning approaches for predicting customer reviews. This study uses three supervised machine learning methods to analyze customer sentiment on eCommerce websites. Details of each machine learning model are given in this study and the overall benefits, capabilities, and performance of each are provided in the context of analyzing customer sentiment (Tribhuvan et al., 2014). The effect of data size and data type and how to get reliable important features and data visualization simultaneously are also evaluated. Lastly, general guidelines are also applied on using and interpreting different supervised machine learning methods for reliable analysis of Amazon Baby Product Review based on its size and complexity (Na et al., 2004).

Objectives

To classify text into positive, negative, or neutral sentiments, which helps in understanding the overall sentiment conveyed by the text and quantifying public opinion or customer feedback.

To extract subjective information and opinions from the text involves identifying the key aspects, features, or entities being discussed and determining the sentiment associated with each aspect.

To analyze sentiments expressed in customer reviews, social media posts, and surveys, sentiment analysis helps businesses gain insights into customer perceptions, satisfaction levels, and preferences. This understanding can guide decision-making and improve products, services, and overall customer experience.

LITERATURE REVIEW

E-commerce is vital for facilitating business interactions between merchants and customers, overcoming geographical barriers. Alongside enabling online product purchases, e-commerce platforms also offer customers the ability to share their reviews and ratings regarding the products (Singh et al., 2013). These opinions hold substantial influence over potential customers who are considering purchasing the same product (Na et al., 2004). Consequently, companies leverage sentiment analysis and opinion-mining techniques on these reviews and ratings to monitor product sales and gauge market value (Angulakshmi et al., 2014). During the analysis phase, sentiment analysis is employed, utilizing a sophisticated lexicon containing predefined positive and negative expressions to determine the sentiment polarity (Li et al., 2013). A literature study explored sentiment analysis within the realms of sentiment dictionaries and machine learning approaches. Determining the semantic orientation of sentences, words, or phrases is crucial to vocabulary-based sentiment analysis. The field of emotion analysis is witnessing significant advancements, with ongoing research dedicated to this area. Sentiment analysis plays a pivotal role in categorizing and classifying sentiments based on textual content. A majority of studies in this field focus on analyzing diverse reviews from various e-commerce platforms, such as customer reviews (positive or negative), media posts, and popular applications. Deep learning techniques have emerged as a favored approach among researchers for studying sentiment analysis. One author conducted a comprehensive sentiment analysis of financial news articles using a combination of lexicon-based and machine learning-dependent methodologies, leading to more reliable findings through nine studies. Achieving accurate text categorization is challenging due to datasets' constantly evolving nature, encompassing various formats, ratings, and blogs (Carenini et al., 2005). To address this challenge, researchers strive to enhance text data classification models. The complex interplay of positive sentiments towards iPhone privacy and negative sentiments regarding iPhone usage adds complexity to iPhone predictions. Sentence-level sentiment analysis is a valuable method for in-depth data analysis, enabling providers and consumers to comprehend crucial aspects that can impact product sales. In a study focused on advertising sharing, the authors proposed the hypothesis that sentiment analysis offers deeper insights into customers' intentions to share online advertisements. Another study introduced an algorithmic approach involving seed opinion lexicons, general word sets, data analysis, and extraction rules for the extraction phase (Tang et al., 2019). As described, learning represents a form of cognitive intelligence. Notably, prior data accompanies research, including feature extraction, collection, validation, and interpretation of testing data, as discussed. The optimization of cloud systems has been explored. Presented a method for emotion analysis utilizing attention-based neural networks and a two-way gated loop unit. Sentiment classification has seen expanded application in recent years, including the use of sentiment analysis techniques to filter feedback on scientific papers. The Twitter text has been subjected to sentiment classification to analyze visitor feedback. Various studies have introduced product-review sentiment analysis applications. The utilization of well-established machine learning (ML) algorithms in sentiment analysis, transforming it into a conventional text classification problem by leveraging syntax and language features. A domain-specific sentiment dictionary has been developed. Twitter sentiment analysis methods, including machine learning techniques and word frequency-based approaches, were discussed. The study compared the efficiency of the Support Vector Machine (SVM) and Naive Bayes (NB) in sentiment analysis, revealing SVM's highest accuracy of 85% when utilizing Bigram models (Govindarajan, 2013). Various methods can be compared to identify the most effective approach for detecting sentiments in Twitter data. The challenges associated with information-related writing, such as achieving equity, relevance, and accountability. They emphasized the role of social networks.

METHODOLOGY

Considerable research has been dedicated to analyzing online social networks, which can be categorized into three main types: geometrical, statistical, and topological. These analysis systems typically involve several steps: detection, extraction, selection, and classification (Rani et al., 2017). These steps effectively determine the network analysis or graph visualization of OSNs, employing various algorithms and techniques. This study also provides a concise overview of social network analysis, emphasizing the visualization of graphs. Visualization techniques for social networks aid in discovering relationships and characteristics among different entities present on these platforms. This study uses Amazon Baby Product reviews to perform sentiment analysis. To accomplish the tasks sequentially, we start by reading the CSV file. In the context of natural language processing (NLP), we begin by removing stop words, which are common words that do not contribute significant meaning to the text (Nasukawa et al., 2003). Calculating the number of stop words can provide additional information about the text that has been omitted.

Furthermore, we explore basic feature extraction techniques from reviews. Before extracting text features, we clean the dataset to ensure the best possible features are obtained. This involves performing preprocessing steps on the training dataset. Figure.1 depicts the phases of the current work starting with the data collection until evaluating each classification model (Figure 1).

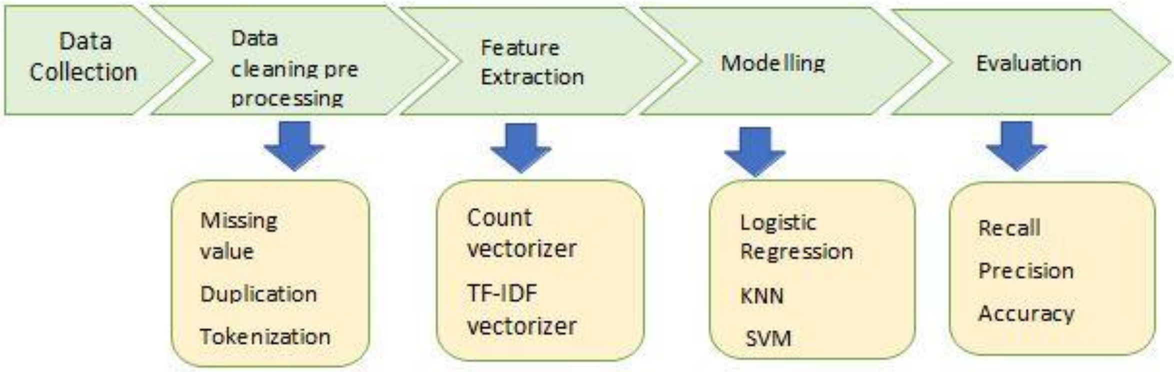


Figure 1
THE OVERALL METHODOLOGY OF SENTIMENT ANALYSIS FOR BABY PRODUCT REVIEWS

Data Collection

A model for better sentiment analysis was designed using an ensemble approach to increase the accuracy and efficiency of reviews for keyword trends. It has been observed minimal variations in ratings between sentence-level and document-level analysis, indicating that sentences can be considered small documents. The dataset for this study was collected from a renowned online platform, "Kaggle." Specifically, a database containing over 175,000 consumer “reviews” and “ratings” of products from Amazon Baby Product Reviews. From this dataset, we selected the second category, consisting of 175,000 user reviews. At amazon.com, reviews are ranked on a scale of 1 to 5. In the context of polarity sentiment classification experiments, ratings of 1 and 2 were considered negative, while ratings of 4 and 5 were regarded as positive. Neutral feedback, denoted by a rating of 3, was also included in the data.

Data Pre-Processing

Data pre-processing is an important step in sentiment analysis to develop the textual data quality; the reviews were pre-processed by altering all the letters to lowercase, not mixed capitals and lowercase. Also, all punctuation and stop words that frequently appear and do not significantly affect meaning were eliminated. All the missing values are replaced. Summary of Pre-processing Steps Each dataset review was labeled positive, negative, or neutral.

Tokenization

Tokenization refers to the process of breaking down a sequence of text into smaller units called tokens. Tokens can be individual words, sentences, or even characters, depending on the specific application and requirements.

Here's a simple example to illustrate tokenization: Input text: "I love playing soccer."

Tokenization: Word tokenization: ["I", "love", "playing", "soccer", "."]

Feature Extraction

Sentiment analysis demands computers to interpret human language. Count Vectorizer and Term frequency - Inverse document frequency (TF-IDF) are being used in this project. Count Vectorizer is used to transform a given text into a vector on the basis of the frequency (count) of each word that occurs in the entire text. TF-IDF will transform the text into a meaningful representation of integers or numbers which is used to fit machine learning algorithms for predictions.

TF-IDF Vectorizer is a measure of the originality of a word by comparing the number of times a word appears in the document with the number of documents the word appears in. formula for TF-IDF is:

TFIDF = TF (t, d) * IDF(t), where, TF = Number of times term t appears in a document d.

IDF = Inverse document frequency.

Classification Models

Classification techniques are applied in the field of Sentiment Analysis to classify data into binary classification (e.g., “positive” and “negative”) and ternary classification (e.g., “positive,” “negative” and “neutral”) and based on that the sentiment analysis process is completed. Two approaches are used in sentiment analysis are following:

Logistic Regression

Logistic regression is used in Sentiment Analysis through ML processes. Its outcome comes under 0 or 1 value. In LR, the regression line looks like a shape form that carries 0 and 1 value (Figure 2).

The curve of the logistic function depicted the likelihood of whether the cells are cancerous or not.

Logistic regression can be used to classify the observations using different types of data and can easily determine the most effective variables for classification. The below image shows the logistic function: The Logistic regression equation can be obtained from the Linear Regression equation. The general form of LR is given below:

$$Y=b_0+b_1x_1+ b_2x_2+\text{--------}+b_nx_n$$

In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y): {(y/(1-y)}; 0 for y=0 and infinity for y=1.

But we need a range between -[infinity] to +[infinity], then take the logarithm of the equation it will become:

$$\text{Log}[y/(1-y)]=b_0+b_1x_1+ b_2x_2+\text{--------}+b_nx_n$$

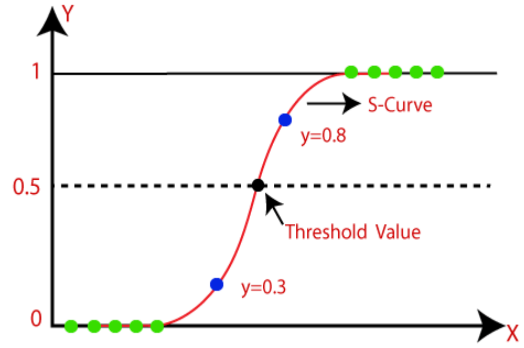


Figure 2

Citation Information: Banerjee K., Sarkar S.,(2024). Sentiment analysis using different machine learning models: A study for the prediction of customer’s review. *Journal of Legal, Ethical and Regulatory Issues, 27*(1), 1-10.

LOGISTIC REGRESSION

K-Nearest Neighbor (KNN) Algorithm

KNN is a supervised learning ML process. KNN algorithm classifies new data based on similarity. It is also called the lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset (Figure 3).

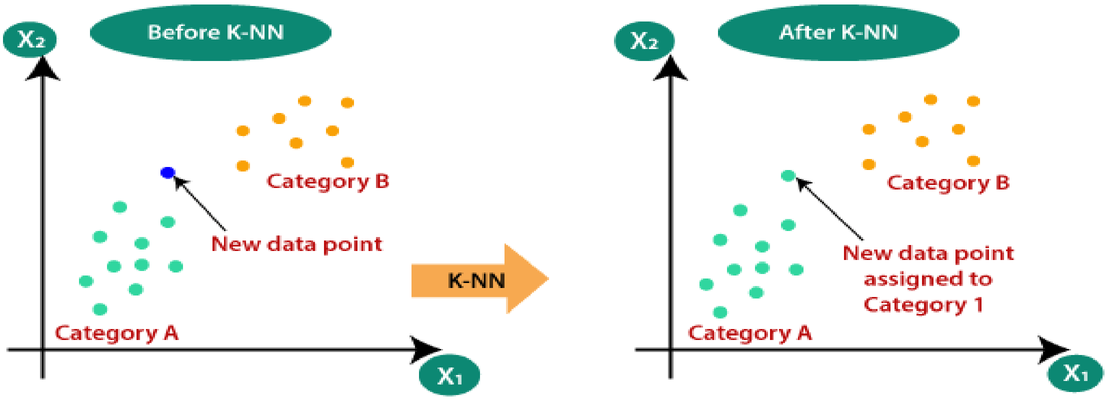


Figure 3
KNN ALGORITHM

Support Vector Machine (SVM) Algorithm

SVM is a supervised learning process in classification. SVMs are particularly useful when the data has many features, and when there is a clear margin of separation in the data. SVM is helpful to perform on multi-class problems, which can create a binary classifier for different data class (Figure 4).

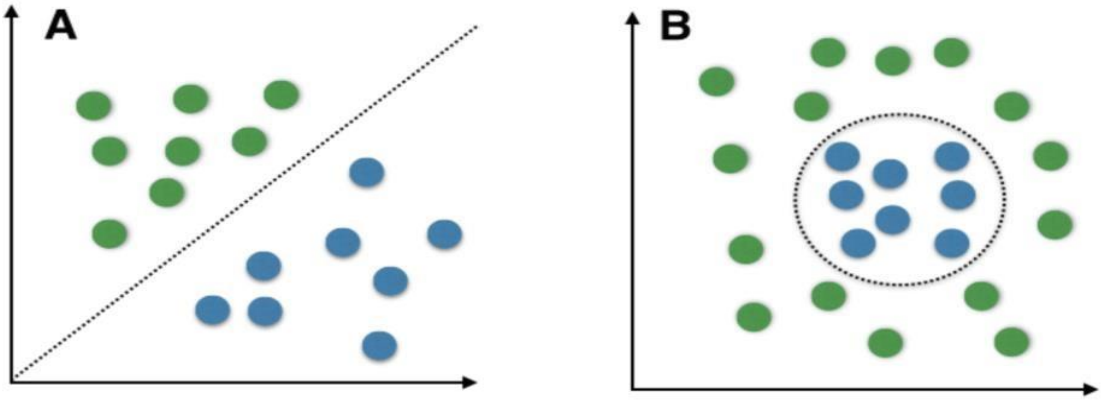


Figure 4
SVM Algorithm

Analysis of Result

Machine learning models play a crucial role in providing valuable predictions for organizations. While training a model is a vital step, it is equally important to assess how well the model performs on new, unseen data. Generalization of the model on such data is essential to determine its reliability and the trustworthiness of its predictions. The evaluation of a model aims to estimate its accuracy in generalizing to future data points that it hasn't encountered during training.

Research has been used in classification problems, namely: Precision and recall, Confusion matrix, and Receiver operating characteristic curve (ROC)

Precision and Recall

Precision measures the proportion of correctly predicted positive instances out of all instances and recall calculates the proportion of correctly predicted positive instances out of all actual positive instances. These metrics help assess the model's ability to accurately identify positive cases and avoid false positives or false negatives (Table 1 & Table 2).

Table 1
PRECISION AND RECALL FOR COUNT VECTORIZER

Model	Precision	Recall
Logistic Regression	0.933	0.955
K Nearest Neighbor	0.859	0.992
Support Vector Machine	1.0	0.818

Source: Author's calculation

Table 2
PRECISION AND RECALL FOR TF-IDF VECTORIZER

Model	Precision	Recall
Logistic Regression	0.912	0.983
K Nearest Neighbor	0.877	0.981
Support Vector Machine	0.987	0.913

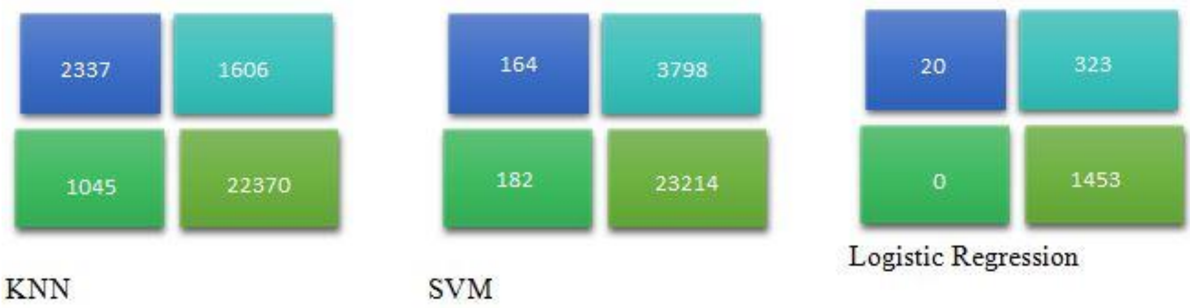
Source: Author's calculation

Confusion Matrix

The confusion matrix presents a tabular representation of the model's predictions versus the actual class labels (Figure 5). It breaks down the predictions into true positives, true negatives, false positives, and false negatives. This matrix offers a comprehensive overview of the model's performance across different classes, enabling further analysis of specific errors or misclassifications (Figure 6 & Figure 7).

True Negative
False Positive
False Negative
True Positive

Figure 5
CONFUSION MATRIX FORMAT



2337	1606
1045	22370

KNN

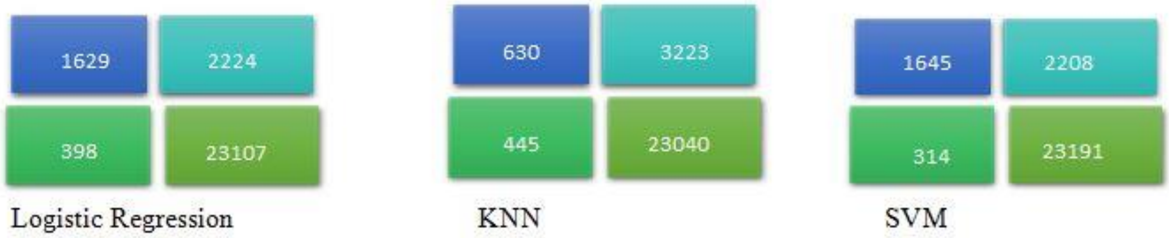
164	3798
182	23214

SVM

20	323
0	1453

Logistic Regression

Figure 6
CONFUSION MATRIX OF COUNT VECTORIZER



1629	2224
398	23107

Logistic Regression

630	3223
445	23040

KNN

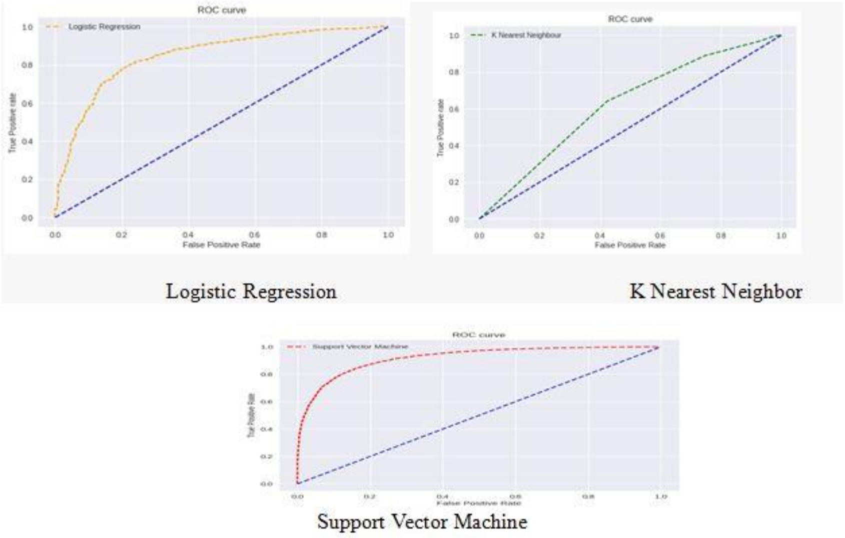
1645	2208
314	23191

SVM

Figure 7
CONFUSION MATRIX OF TF-IDF VECTORIZER

Receiver Operating Characteristic Curve (ROC)

The ROC curve is a graphical representation of the model's performance by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) at various classification thresholds. It provides insights into the trade-off between true positive rate and false positive rate and helps determine an optimal threshold for classification (Figure 8 & Figure 9).

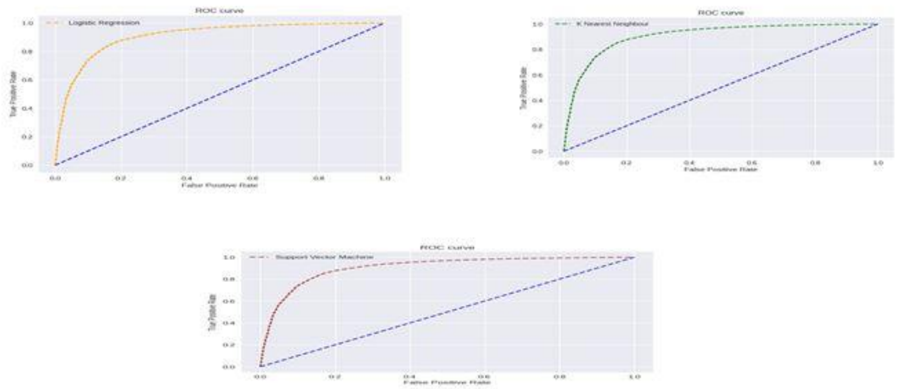


Logistic Regression

K Nearest Neighbor

Support Vector Machine

Figure 8
ROC CURVE OF CLASSIFICATION MODELS FOR COUNT VECTORIZER



Support Vector Machine

Figure 9
ROC CURVE OF CLASSIFICATION MODELS FOR TF-IDF VECTORIZER

Results obtained using Train-Test split

Table 3
ACCURACY AND ROC-AUC FOR COUNT VECTORIZER

Model	Accuracy	ROC-AUC
Logistic Regression	0.90	90.74
K-Nearest Neighbor	0.86	63.90
Support Vector Machine	0.82	83.00

Source: Author's calculation

Table 4
ACCURACY AND ROC-AUC FOR TF-IDF VECTORIZER

Model	Accuracy	ROC-AUC
Logistic Regression	0.90	50.20
K-Nearest Neighbor	0.87	93.60
Support Vector Machine	0.91	96.30

Source: Author's calculation

From the above analysis research is going to highlights the following major observations (Table 3 & Table 4):

a) TF-IDF vectorizer gave us better results than the Count vectorizer in Support Vector Machine Model in the K Nearest Neighbor model and Logistic Regression. The Count vectorizer gave us better results in Logistic Regression than the KNN and SVM model. TF-IDF vectorizer in SVM gives better results than the Count vectorizer in Logistic regression.

b) KNN was the best out of these two classifiers used for this project considering overall accuracy, true positive rate, and true negative rate.

c) From the above ROC curve, we can conclude that the Support Vector Machine algorithm using the TF-IDF vectorizer model best fits this dataset.

d) We will use a TF-IDF vectorizer in Support Vector Machine Model to get the best possible accuracy.

CONCLUSION

Sentiment Analysis is a valuable tool for understanding the opinions and attitudes expressed in text data. Using machine learning algorithms, sentiment analysis can accurately classify text into positive, negative, and neutral categories. We implemented various sentiment analysis models during the project and evaluated their performance on a given dataset. We experimented with different feature extraction techniques such as count and TF-IDF vectorizer and explored different classification algorithms, including logistic regression, and K Nearest Neighbor. And Support Vector Machine. The results from the study showed that in terms of accuracy, the KNN approach achieves better results than the Logistic Regression approach when the whole data set was used as a training and testing data set. Overall, the project demonstrated the potential of sentiment analysis in customer feedback analysis. Sentiment analysis will continue to improve and provide valuable insights for businesses and organizations.

REFERENCES

Agarwal, A., & Sabharwal, J. (2012, December). End-to-end sentiment analysis of twitter data. In *Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data* (pp. 39-44).

Angulakshmi, G., & ManickaChezian, R. (2014). An analysis on opinion mining: techniques and tools. International Journal of Advanced Research in Computer and Communication Engineering, 3(7), 2319-5940.

Cambria, E., Poria, S., Gelbukh, A., & Thelwall, M. (2017). Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, *32*(6), 74-80.

Carenini, G., Ng, R. T., & Zwart, E. (2005, October). Extracting knowledge from evaluative text. In *Proceedings of the 3rd international conference on Knowledge capture* (pp. 11-18).

Chandrakala, S., & Sindhu, C. (2012). Opinion mining and sentiment classification: A survey. *ICTACT journal on soft computing*, *3*(1), 420-425.

Ingle, A., Kante, A., Samak, S., & Kumari, A. (2015). Sentiment analysis of twitter data using hadoop. *International Journal of Engineering Research and General Science*, *3*(6), 144-147.

Li, S., Wang, Z., Lee, S. Y. M., & Huang, C. R. (2013, August). Sentiment classification with polarity shifting detection. In *2013 International conference on Asian language processing* (pp. 129-132). IEEE.

M.Govindarajan (2013), Sentiment Analysis of Movie Reviews using Hybrid Method of Naive Bayes and Genetic Algorithm, International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970), Volume-3 Number-4 Issue-13, pp 45-67

Na, J. C., Sui, H., Khoo, C., Chan, S., & Zhou, Y. (2004). Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews. *Advances in Knowledge Organization*, *9*, 49-54.

Nasukawa, T., & Yi, J. (2003, October). Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture* (pp. 70-77).

Rani, S., & Kumar, P. (2017). A sentiment analysis system to improve teaching and learning. *Computer*, *50*(5), 36-43.

Sahayak, V., Shete, V., & Pathan, A. (2015). Sentiment analysis on twitter data. International Journal of Innovative Research in Advanced Engineering (IJIRAE), 2(1), 178-183.

Singh, V. K., Piryani, R., Uddin, A., & Waila, P. (2013, March). Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. In 2013 International mutli-conference on automation, computing, communication, control and compressed sensing (imac4s) (pp. 712-717). IEEE.

Tang, D., Qin, B., & Liu, T. (2015). Deep learning for sentiment analysis: successful approaches and future challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *5*(6), 292-303.

Tribhuvan, P. P., Bhirud, S. G., & Tribhuvan, A. P. (2014). A peer review of feature based opinion mining and summarization. *IJCSIT) International Journal of Computer Science and Information Technologies*, *5*(1), 247-250.

Received: 02-Nov-2023, Manuscript No. JLERI-23-14257; **Editor assigned:** 04-Nov-2023, Pre QC No. JLERI-23-14257(PQ); **Reviewed:** 18-Nov-2023, QC No. JLERI-23-14257; **Revised:** 23-Nov-2023, Manuscript No. JLERI-23-14257(R); **Published:** 30-Nov-2023