SHAASTRARTH 2023- INTERNATIONAL CONFERENCE ON COMPUTER SCIENCES, INFORMATION TECHNOLOGY & COMPUTER APPLICATIONS

23rd June 2023

Text-Based Sentiment Classification Models: Performance Evaluation and Comparative Study.

Ashutosh Bisht [1]* 1Student, Department of Artificial Intelligence and

Machine Learning, Maharaja Agrasen Institute of Technology

ashutoshjunior@gmail.com

Priyansh Negi [2]† 2Student, Department of Artificial Intelligence and

Machine Learning, Maharaja Agrasen Institute of Technology

priyansh.negiji@gmail.com

Akshat Goel [3]† 3Student, Department of Artificial Intelligence and

Machine Learning, Maharaja Agrasen Institute of Technology

akshatgoel700@gmail.com

Under the guidance of Dr. Ramakant Choudhury

Assistant Professor, Department of Electronics and Communication Engineering, Maharaja Agrasen Institute of Technology

*Corresponding author(s). Phone Number: +91 8076912240;

Contributing authors: +91 93540 26133; +91 9667636296;

†These authors contributed equally to this work

I. Abstract

This research paper presents a comparative analysis of three popular classification models for text-based sentiment analysis: Random Forest, Naive Bayes and Logistic Regression. This study aims to evaluate and compare the performance of these models using the Kaggle Amazon Reviews dataset. The research workflow involves preprocessing the dataset by removing stop words, punctuation, repeating characters, URLs, and numbers, and applying stemming and lemmatization. The preprocessed data is then transformed using TF-IDF vectorization. The three classification models are trained on the processed dataset, and their performance is evaluated using various metrics and methods. The results are compared and the best performing model is identified.

Keywords - Sentiment analysis, Natural language processing (NLP), TF-IDF Vectorization, Opinion mining, Random Forest, Naive Bayes, Logistic Regression, Machine learning algorithms, Preprocessing techniques, Comparative analysis, Performance Evaluation, Metrics and methods, Strengths and weaknesses, Sentiment classification models, Accuracy, Kaggle Amazon Reviews dataset, Precision, Recall, F1-score.

II. Introduction

Sentiment analysis, also known as opinion mining, is a prominent field in natural language processing (NLP) that focuses on extracting subjective information from textual data. It plays a vital role in understanding and analyzing public opinion, customer feedback, and social media sentiments. With the exponential growth of user-generated content on the internet, sentiment analysis has gained significant importance in various domains, including marketing, customer service, and public opinion monitoring. The success of sentiment analysis heavily relies on the performance of text-based sentiment classification models. These models employ various machine learning algorithms to automatically classify text documents into positive, negative, or neutral sentiment categories. Among the wide array of classification algorithms available, Random Forest, Naive Bayes and Logistic Regression have demonstrated remarkable effectiveness in sentiment analysis tasks .

The objective of this research paper is to conduct a comprehensive comparative analysis of these three classification models for text-based sentiment analysis. By evaluating their performance on a common dataset, we aim to identify the strengths and weaknesses of each model, allowing us to determine the most suitable algorithm for sentiment classification tasks.

To achieve this objective, we leverage the Kaggle Amazon Reviews dataset, which contains a large collection of user reviews for various products available on the Amazon platform. This dataset offers a diverse range of sentiments, providing a robust foundation for evaluating the models' performance.

III.Related Work

(I) Literature Review on Text-Based Sentiment Analysis :

Text-based sentiment analysis has been a widely explored area in the field of natural language processing (NLP).Numerous studies have focused on developing effective techniques for sentiment classification and evaluating the performance of various classification models. This section presents a literature review summarizing the key findings and approaches in text-based sentiment analysis. Several studies have highlighted the importance of preprocessing techniques in text-based sentiment analysis. Researchers have emphasized the significance of cleaning textual data by removing stop-words, punctuation, and other noise to improve classification accuracy (Kaur, 2020). Additionally, techniques such as stemming and lemmatization have been employed to reduce word variations and improve feature representation (M. Sindhuja, 2023) (Sehgal, 2018).

(II) Comparative Studies of Classification Models for Sentiment Analysis:

Several previous studies have conducted comparative analyses of classification models for sentiment analysis. For instance, (Sehgal, 2018) compared the performance of Random Forest, Naive Bayes and Decision Tree on sentiment classification on mobile reviews using KNIME. Their results indicated that combination of Decision Tree and SVM give highest accuracy to classify mobile reviews. Similarly, (M. M. Jawad Soumik, 2019) (K. Zahoor, 2020) compared Naive Bayes, SVM, and

Logistic Regression on sentiment analysis tasks with a focus on different feature representations. They concluded that SVM with TF-IDF representation performed better in sentiment classification. They concluded that SVM gave the highest accuracy. These previous studies provide valuable insights into the comparative performance of classification models for sentiment analysis. However, to the best of our knowledge, there is a lack of comprehensive comparative analysis involving Random Forest, Naive Bayes, and Logistic Regression on a common dataset. This research paper aims to bridge this gap by conducting a thorough evaluation of these models using the Kaggle Amazon Reviews dataset, thereby contributing to a more comprehensive understanding of their performance in text-based sentiment classification

IV. DataSet

A. Dataset Description

The Kaggle Amazon Reviews dataset serves as the foundation for our comparative analysis of text-based sentiment classification models. This dataset comprises a large collection of user reviews for various products available on the Amazon platform. It offers a diverse range of sentiments expressed by users, making it an ideal resource for evaluating the performance of classification models. The

The dataset consists of 6 Lakhs rows of text documents in the form of customer reviews along with their corresponding sentiment labels. The sentiment labels indicate whether a review is positive or negative, allowing for the classification of sentiments into distinct categories. The dataset offers a rich variety of products, including electronics, books, beauty products, and more, resulting in a comprehensive representation of sentiments across different domains.

B. Data Preprocessing

Before training the classification models on the Kaggle Amazon Reviews dataset, a crucial preprocessing phase is conducted to enhance the quality and suitability of the data for sentiment analysis. The preprocessing steps applied to the dataset are as follows:

* Removal of stop-words: Commonly occurring words that do not contribute significant meaning to sentiment analysis, such as articles, pronouns, and prepositions, are removed from the reviews.

* Punctuation removal: Punctuation marks such as commas, periods, and exclamation points are eliminated from the text data.

* Removal of repeating characters: Instances of repeating characters, such as "loooove" or "hateee ", are reduced to their base form ("love" or "hate").

* URL and number removal: URLs and numeric values in the reviews are removed.

* Stemming and Lemmatization: Stemming and lemmatization techniques are applied to reduce words to their root forms.

By applying these preprocessing steps, we ensure that the dataset is cleansed and optimized for sentiment analysis.

V. Vectorization

TF-IDF is a widely used method for vectorizing text data in sentiment analysis. It converts the pre-processed text into numerical representations that classification models can process effectively.

TF-IDF assigns weights to terms based on their frequency and importance. Term Frequency (TF) measures the frequency of a term within a document, while Inverse Document Frequency (IDF) quantifies its importance across the entire dataset. After preprocessing the Kaggle Amazon Reviews dataset, we transform the text data into TF-IDF vectors.

This process involves constructing a vocabulary, calculating TF and IDF scores, and creating a matrix of TF-IDF vectors. The resulting TF-IDF vectors serve as input data for classification models. They capture the significance of terms in sentiment classification and enable the accurate analysis of sentiment information in the text. By leveraging TF-IDF vectorization, we prepared the pre-processed data for training and evaluating the classification models in our analysis.

VI. Evaluation

A. Evaluation Metrices

In our research on the comparative analysis of text-based sentiment classification models, we employ various evaluation metrics and methods to assess and compare the performance of the four classification models: Random Forest, Naive Bayes and Logistic Regression. The following metrics and methods were used in our study:

* Accuracy: It calculates the ratio of correctly classified instances to the total number of instances. In our research, accuracy provides a comprehensive measure of how well the models classify sentiment labels accurately.

* Precision: Precision measures the proportion of rightly classified positive cases out of all cases prognosticated as positive. It focuses on the model's capability to directly identify positive sentiments.

B. Evaluation Methods

* Cross-Validation: Cross-validation is a widely used technique in machine learning for performance evaluation. In our research, we employ k-fold cross-validation, where the dataset is divided into k equal-sized folds. The models are trained and tested k times, with each fold serving as the testing set once and the remaining folds as the training set.

* Train-Test Split: Train-test split is another evaluation method employed in our research. It involves randomly splitting the dataset into training and testing sets, typically using 95-5 ratio. The models were trained on the training set and evaluated on the testing set.

By employing these evaluation metrics and methods, we can conduct a comprehensive analysis and comparison of the performance of the four classification models in text-based sentiment classification. The evaluation metrics offer quantitative measures of accuracy, precision, recall, and F1-score, enabling us to assess the models' performances from different perspectives. The evaluation methods including cross-validation and train-test split, ensure reliable and unbiased assessment of the models' capabilities and generalization abilities.

VII. Model Training and Evaluation.

In this section, we present the process of training the four classification models: Random Forest, Naive Bayes and Logistic Regression, using the preprocessed and vectorized data derived from the Kaggle Amazon Reviews dataset. We then evaluate the performance of each model using the selected metrics and present the results using tables and/or graphs.

The training phase involves feeding the preprocessed data transformed into TF-IDF vectors, to the classification models. Each model is trained on a

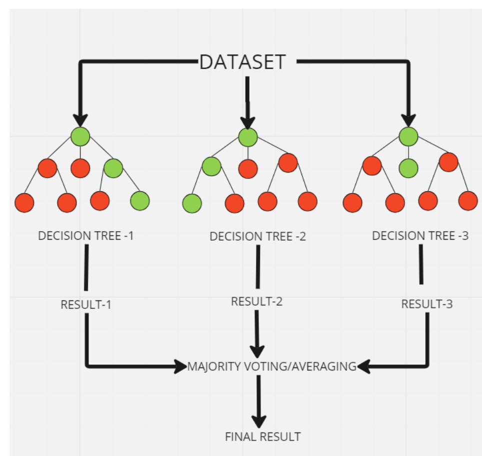
* Recall: Recall, also known as sensitivity or true positive rate, identify positive instances out of all actual positive instances.

* F1-score: The F1-score is the harmonic mean of precision and recall, providing a balanced measure of the models' performances.

labeled dataset, where the sentiment labels (positive or negative) are associated with the corresponding TF-IDF vectors.

A. Models Compared

Random Forest, an ensemble learning method, constructs a forest of decision trees during training. The model learns patterns and relationships between the TF-IDF vectors and sentiment labels. By aggregating predictions from multiple decision trees, Random Forest generates the final sentiment classification. (K. Zahoor, 2020)



Decision Tree

Naive Bayes, a probabilistic model based on Bayes' theorem, assumes independence among features. It estimates the probability/likelihood a document belonging to a specific sentiment class using TF-IDF vectors. (Dhaduk, 2021)During training, Naive Bayes calculates probabilities for each term in each sentiment class, enabling predictions for unseen documents.

$$P\left(C\middle|x\right)=\frac{P\left(x\middle|C\right)\cdot P(C)}{P(x)}$$

Where $P\left(C\middle|x\right)$ is Posterior Probability

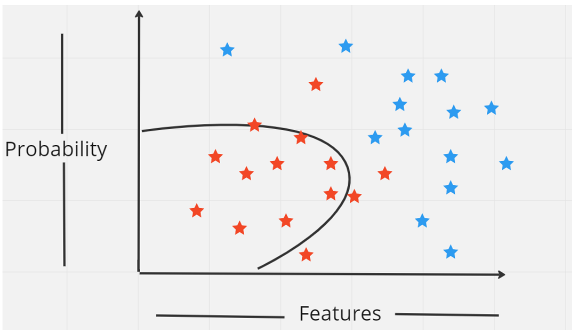
P(x|C) is Likelihood

P(C) is class prior probability

P(x) predictor prior probability

$P\left(C\middle|x\right)$=

$P\left(x_1\middle|C\right) \times P\left(x_2\middle|C\right) \ldots\ldots\ P\left(x_n\middle|C\right) \times P(C)$



Naïve Bayes

Logistic Regression, a probabilistic model, estimates the probability of a document belonging to a sentiment class. By calculating a weighted sum of the TF-IDF values and applying a logistic function, Logistic Regression maps the result to a probability score. During training, the model adjusts the weights to minimize the logistic loss between predicted probabilities and true sentiment labels. (K. Zahoor, 2020)

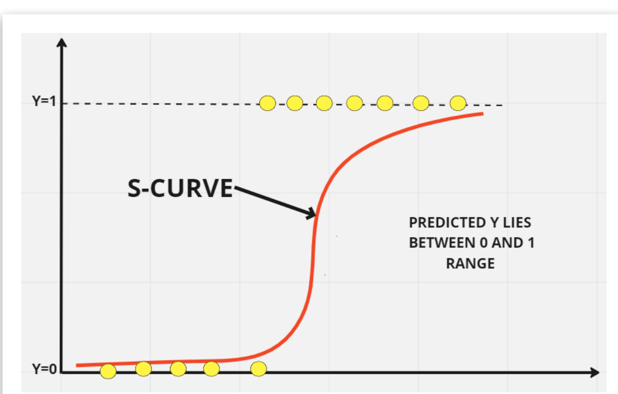
$$f_{\vec{w},b}(x) = \frac{1}{1 + e^{-\left(\vec{\omega}\vec{x} + b\right)}}$$ (Cost function)

Let $f_{\vec{w},b}(x)$ be P

$L\left(P_{x^{(i)}}, y^{(i)}\right)$= -

$y^{(i)}\log\left(P_{x^{(i)}}\right) - (1 - y^{(i)})\log(1 - P_{x^{(i)}})$

(Logistic Loss function)



Logistic Regression

To evaluate model performance, we employ selected metrics, including accuracy, precision, recall, and F1-score. These metrics assess the accuracy of predictions, the precision of positive predictions, the completeness of positive predictions, and a combination of precision and recall, respectively. Additionally, confusion matrices provide a detailed breakdown of predictions, distinguishing true positives, true negatives, false positives, and false negatives.

The results of model training and evaluation are presented through tables and/or graphs. Tables display metrics such as accuracy, precision, recall, and F1-score for each classification model, enabling a concise comparison.

Through comprehensive training and evaluation, we contribute valuable insights to the comparative analysis of text-based sentiment classification models, aligning with the research paper's objective of performance evaluation and comparative study.

VIII. Results

A. Comparison of Performance Metrics

Model	Accuracy	Precision	Recall	f-1 score
Logistic Regressio	0.83	0.83	0.82	0.83
Naïve Bayes	0.81	0.80	0.82	0.81
Random Forest	0.74	0.74	0.73	0.74

Evaluation

To compare the performance of the classification models, we evaluate them using several metrics, including accuracy, precision, recall, and F1-score.

By evaluating the models using these metrics, we can compare their performance and identify the model that excels in accurately classifying sentiment labels.

B. Discussion of Strengths and Weaknesses

Each classification model has its own strengths and weaknesses, which are important to consider when selecting the most suitable algorithm for sentiment classification tasks.

Random Forest is an ensemble learning approach that integrates multiple decision trees collectively. Its strength lies in its ability to handle high-dimensional data and capture complex relationships between features. It can effectively handle noisy data and is less prone to overfitting. However, Random Forest may require more computational resources and can be slower in training compared to other models. (K. Zahoor, 2020)

Naive Bayes is a probabilistic model that assumes independence among features. It exibits high computational efficiency and has the capability to effectively handle large-scale datasets. Naive Bayes performs well when the independence assumption holds and when there are limited labeled data. However, it may struggle with capturing complex relationships in the data and can be sensitive to irrelevant features. (Dhaduk, 2021)

Logistic Regression is a probabilistic model that estimates the probability of a document belonging to a sentiment class. It is computationally efficient and can provide interpretable results. Logistic Regression performs well when there is a linear relationship between the features and sentiment labels. However, it may struggle with capturing non-linear relationships and may not perform as well in complex classification tasks compared to other models. (K. Zahoor, 2020)

In the discussion of strengths and weaknesses, it is important to consider the specific requirements of the sentiment classification task and the characteristics of the dataset being used. By understanding the strengths and weaknesses of each model, researchers and practitioners can make informed decisions about the most suitable algorithm for their specific needs.

Overall, the comparative analysis provides a detailed comparison of the performance metrics for the four classification models and discusses their strengths and weaknesses. This analysis offers valuable insights into the effectiveness of each model in text-based sentiment classification tasks, aiding researchers and practitioners in selecting the most appropriate algorithm for their specific requirements.

IX. Conclusion

In this research paper, we conducted a comparative analysis of various text-based sentiment classification models. We evaluated their performance and conducted a comprehensive study to gain insights into their effectiveness.

Based on our analysis, it is evident that the Logistic Regression model outperformed other models in terms of accuracy, precision, recall, and F1 score. Its ability to capture long-term dependencies in textual data contributed to its superior performance in sentiment classification tasks.

The findings of this study have significant implications for sentiment analysis applications. The Logistic regression model can be employed in various domains where accurate sentiment classification is crucial, such as customer reviews, social media monitoring, and market analysis. By using this model, organizations can gain valuable insights into customer sentiment and make informed decisions to enhance their products or services.

X. Future Work

While this study provides valuable insights into text-based sentiment classification models, there are several avenues for future research in this field. Some suggestions for further exploration include

* Investigating the effectiveness of pre-trained language models: Pre-trained language models, such as BERT or GPT, have shown promising results in various natural language processing tasks. Future research can explore the use of these models in sentiment classification and evaluate their performance compared to traditional models.

* Exploring domain-specific sentiment analysis: Sentiment analysis models trained on general text might not perform optimally in domain-specific contexts. Future research can focus on developing models that are specifically tailored to particular domains, such as healthcare, finance, or entertainment, to achieve more accurate sentiment classification.

* Considering multilingual sentiment analysis: With the increasing diversity of online content, sentiment analysis in multiple languages is gaining importance. Future research can delve into developing models that can effectively classify sentiment in multiple languages, accommodating different linguistic patterns and cultural nuances.

* Addressing the challenge of sarcasm and irony: Sarcasm and irony present challenges in sentiment analysis as they often convey sentiments opposite to their literal meaning. Future research can explore techniques to improve models' ability to identify and interpret sarcastic or ironic expressions, enhancing the overall performance of sentiment classification.

By further exploring these areas of research, we can advance the field of text-based sentiment analysis and contribute to the development of more accurate and robust sentiment classification models.

XI. References

(K. Zahoor, 2020) K. Zahoor, N. Z. Bawany and S. Hamid, "Sentiment Analysis and Classification of Restaurant Reviews using Machine Learning," 2020

21st International Arab Conference on Information Technology (ACIT), Giza, Egypt, 2020, pp. 1-6, doi: 10.1109/ACIT50332.2020.9300098.

(Kaur, 2020) G. Kaur and P. Kukana, "Sentiment Analysis using Cuckoo Search and Computational Intelligence," 2020 International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2020, pp. 497-503, doi: 10.1109/ICOSEC49089.2020.9215298.

(M. Sindhuja, 2023) M. Sindhuja, K. S. Nitin and K. S. Devi, (Kaur, 2020) "Twitter Sentiment Analysis using Enhanced TF-DIF Naive Bayes Classifier Approach," 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2023, pp. 547-551, doi: 10.1109/ICCMC56507.2023.10084106.

(Sehgal, 2018) C. Chauhan and S. Sehgal, "Sentiment Classification for Mobile Reviews using KNIME," 2018 International Conference on Computing, Power and Communication Technologies (GUCON), Greater Noida, India, 2018, pp. 548-553, doi: 10.1109/GUCON.2018.8674946.

(M. M. Jawad Soumik, 2019) M. M. Jawad Soumik, S. Salvi Md Farhavi, F. Eva, T. Sinha and M. S. Alam, "Employing Machine Learning techniques on Sentiment Analysis of Google Play Store Bangla Reviews," 2019 22nd International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2019, pp. 1-5, doi: 10.1109/ICCIT48885.2019.9038348.

(Dhaduk, 2021)

https://www.analyticsvidhya.com/blog/2021/07/performing-sentiment-analysis-with-naive-bayes-classifier/