Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Proposed Algorithm and Models for Sentiment Analysis and Opinion Mining Using Web Data

[1]Nadiya Parveen, [2]Mohd Waris Khan,
[1]Research Scholar, Department of Computer Application, Integral University, Lucknow, India, 786nadiyaparveen@gmail.com
[2]Assistant Professor, Department of Computer Application, Integral University, Lucknow, India, wariskhan070@gmail.com

Abstract
Individuals are increasingly expressing their ideas and emotions on social media platforms due to the continuous rise in content publication. With this in mind, the use of sentiment analysis and opinion mining techniques might prove to be of tremendous assistance in the process of automated textual corpus analysis (which includes things like comments, reviews, tweets, and other similar content). Despite the significant progress in text mining algorithms, deep learning approaches, and text representation models, the results for high-density languages like English, possessing vast training corpora and a wealth of linguistic resources, remain relatively robust. Nevertheless, there is still opportunity for development in relation to activities that include languages with a lower density. This work investigates the construction of an algorithm to evaluate the correctness of models and the influence that various data preparation approaches, such as stemming or deleting stop words, have on the performance of machine learning models. When it comes to monitoring feelings, opinion mining and sentiment analysis are two of the most important technologies. Additionally, in order to do a sentiment analysis for reviews that we have gathered from Twitter and Flipkart's, we compare a number of different categorization approaches and learning methods by making use of the provided algorithms. This model can use any information from Facebook, Twitter, or any other microblogging site.
Index terms Sentiment Analysis, Opinion Mining, Web Data, Natural Language Processing, Machine Learning

I Introduction

Social media is now the largest source of public opinion on a wide range of topics, including people, businesses, locations, and ideas. This is mostly because individuals want to use social media to remark on other people's thoughts and share their own perspectives on daily, local, or global concerns. This leads to the creation of a vast quantity of opinionated data. Marketing firms that create campaigns and promote goods, people, and ideas may find useful information about interests, attitudes, trends, dangers, and possibilities via the appropriate administration and analysis of such data. Sentiment analysis tasks include subjectivity and opinion polarity mining, binary or multi-class emotion classification, and emotion detection (Georgios Paltoglou, 2016, Antonio Reyes, 2013). More in-depth descriptions of public opinion may also be obtained via the use of argument extraction and opinion summarizing(V. Bobichev et. al., 2017). Several granularities of analysis are possible: from the mood of the whole text or of individual sentences to the opinion of a particular entity or facet. To increase outcomes, some methods include a large set of methodologies, such as knowledge bases (dictionaries, corpora, and lexicons), together with unsupervised learning. The exponential growth of web data, fueled by the proliferation of social media, online forums, blogs, and review sites, has transformed the internet into a vast repository of public opinion. This data encompasses a wide range of topics, from consumer preferences and political views to sentiments about current events and product feedback. Extracting meaningful insights from this massive volume of unstructured text presents a significant challenge and an unparalleled opportunity.Sentiment analysis and opinion mining have emerged as critical tools for interpreting and leveraging this data**(A. Shevtsov et. al, 2022).** These fields focus on determining the sentiment expressed in text and extracting subjective

information, respectively. By understanding the emotional tone and opinions conveyed in web data, businesses, policymakers, and researchers can make informed decisions and tailor their strategies to align with public sentiment.Traditional approaches to sentiment analysis and opinion mining have relied heavily on supervised learning techniques, requiring extensive labelled datasets for training (Pramod D, 2022). However, the dynamic and diverse nature of web data necessitates more sophisticated methods. Advanced algorithms that integrate semi-supervised and unsupervised learning, along with comprehensive knowledge bases such as dictionaries, corpora, and lexicons, have shown promise in enhancing the accuracy and depth of sentiment analysis. This research aims to develop and refine advanced algorithms for sentiment analysis and opinion mining, leveraging the vast amounts of web data available. By employing state-of-the-art machine learning techniques and incorporating linguistic resources, we seek to create robust models capable of capturing nuanced sentiments and opinions across various domains (Sutoyo E, 2022, Tao X, 2016, Zhu J.J., Chang, 2016). The goal is to not only improve the precision of sentiment detection but also to facilitate a deeper understanding of the underlying attitudes and trends that shape public discourse(E. Chen, 2022).The outcomes of this research will have wide-ranging applications, and research. As we continue to navigate an increasingly interconnected and opinionated digital landscape, the development of advanced sentiment analysis and opinion mining algorithms will play a pivotal role in harnessing the power of web data for societal benefit.The development of advanced algorithms for sentiment analysis and opinion mining using web data represents a significant step forward in understanding and leveraging public opinion. By integrating deep learning models, transfer learning, and linguistic resources, researchers can create robust models capable of capturing nuanced sentiments across various domains. The applications of these advancements are vast, spanning marketing, politics, finance, healthcare, and beyond.As we continue to traverse a digital universe that is becoming more linked and opinionated, the significance of sentiment analysis and opinion mining will only continue to increase (R. Chandra et al., 2022). Addressing the challenges and ethical considerations associated with these fields will be crucial for harnessing the power of web data for societal benefit. The ongoing research and development in this area promise to unlock new insights and opportunities, shaping the future of how we understand and interact with public opinion.

IISentiment Lexicon and Its Issues

A sentiment lexicon can be defined as a curated collection of words and phrases associated with specific sentiment values. Positive, negative, or neutral attitudes often categorize these values. This instrumental tool, which serves as a core component of sentiment analysis, facilitates the detection and measurement of sentiment in textual data.However, sentiment lexicons face several challenges and limitations. One major issue is the context-dependence of sentiment; words can convey different sentiments depending on their usage and surrounding context, leading to inaccuracies in sentiment classification**(Susmitha et. al., 2024)**Lexicons often struggle with handling slang, idioms, and evolving language, particularly prevalent in social media and informal communication**(**P. Wicke et. al**, 2021).** The lack of domain-specificity is another problem, as a general sentiment lexicon may not accurately capture sentiments in specialized fields like finance, healthcare, or technology. Furthermore, multilingual sentiment analysis poses a significant challenge due to the need for comprehensive lexicons in various languages, which are often underdeveloped or unavailable. Despite these issues, sentiment lexicons remain a valuable resource in the field of sentiment analysis, though they are increasingly supplemented with advanced machine learning techniques to improve accuracy and adaptability.A sentiment lexicon is a curated list of words and phrases associated with specific sentiment values, typically classified as positive, negative, or neutral (Sánchez-Núñez, 2020). It serves as a fundamental tool in sentiment analysis, aiding in the identification and quantification of sentiment in textual data. For example, a simple sentiment lexicon might classify the word "happy" as positive, "sad" as negative, and "neutral" as neutral. Sentiment lexicons face several challenges and limitations. One major issue is the context-dependence of sentiment; words can convey different sentiments depending on their usage and surrounding context, leading to inaccuracies in sentiment classification. For instance, the word "happy" is generally positive, but in the phrase "I'm not happy with the service," it expresses a negative sentiment. Lexicons often struggle with handling slang, idioms, and evolving language, particularly prevalent in social media and informal communication**(**Kumar, N. P

et. al.**, 2023).** For example, the slang term "sick" can mean "ill" (negative) or "cool" (positive) depending on the context, which a simple lexicon might not accurately capture.The lack of domain-specificity is another problem, as a general sentiment lexicon may not accurately capture sentiments in specialized fields like finance, healthcare, or technology. For example, the word "growth" is positive in a business context but can be negative in a medical context (e.g., tumor growth). Multilingual sentiment analysis poses a significant challenge due to the need for comprehensive lexicons in various languages, which are often underdeveloped or unavailable (Vidisha M. Pradhan, 2016). For instance, a sentiment lexicon for English might not be directly translatable to Hindi, where cultural nuances and word usage differ significantly.Despite these issues, sentiment lexicons remain a valuable resource in the field of sentiment analysis, though they are increasingly supplemented with advanced machine learning techniques to improve accuracy and adaptability. For example, machine learning models can learn from context and handle the nuances of language, slang, and idioms better than traditional lexicon-based methods, thereby enhancing sentiment analysis performance.

IIIProposed Algorithms

An algorithm to perform sentiment analysis using machine learning with different feature selection and feature extraction techniques to optimize predictive performance.

Input:

- Training dataset with corresponding labels.
- Parameters for Data Splitting i.e., test data size, random state.

Output:

- Model Performance Metrics: Accuracy Scores: Reflects how well each model predicts the target variable on the test set, with higher scores indicating better predictive performance.

Algorithm:

Step 1: Data Preprocessing:

1.1: Tokenizing text, deleting stop words, and converting text data to lowercase are all examples of this.

1.2: Perform lemmatization on the text data.

Step 2: Data Splitting

2.1: Extract the input features (X) and target variable (y) from the DataFrame.

Step 3: Bag of Words

3.1: Convert text data (X) into numerical data using the Count Vectorizer technique.

Step 3: Model Evaluation with All Features:

Step 3.1: Split the features into training and testing sets.

Step 3.2:

Do

For each model i.e., Logistic Regression, Decision Tree, Multinomial NB:

- Training the model using the training set is the next step.
- Predict the results of the test set.
- Determine the result of the calculation and print it out.

End For

Step 4: Model Evaluation with ANOVA Feature Selection Techniques:

Step 4.1: Apply the ANOVA feature selection technique to select the ‘k’ best features.

Step 4.2: Split the selected features into training and testing sets.

Step 4.3: Repeat the step 3.2

Step 5: Model Evaluation with ANOVA and Linear Discriminant Analysis:

Step 5.1: Make use of the 'train_test_split' function in order to divide the ANOVA-selected feature and target variable into training and testing sets with a test size of twenty percent.

Step 5.2: Initialize the Linear Discriminant Analysis (LDA) model.

Step 5.3: Fit the LDA model on the training data and transform both the training and testing datasets using the fitted LDA model.

Step 5.4: Repeat the step 3.2

Step 6: Model Performance

6.1: Select the model and feature selection technique with highest accuracy.

End of Algorithm

IV Methodology

In our methodology, we leveraged publicly available data from two open platforms: Flipkart and Twitter, to conduct a comprehensive sentiment analysis and develop a sophisticated machine learning model for sentiment classification. The process began with data collection, where we extracted a substantial amount of user reviews from Flipkart and tweets from Twitter. These datasets, accessible through open platforms, provided a rich and diverse source of opinions and sentiments related to various products and topics, forming a solid foundation for training an effective sentiment analysis model.The first step involved extensive data preprocessing to ensure data quality and consistency. This included cleaning the text by removing special characters, URLs, and stop words, which could introduce noise into the analysis. Once the data was pre-processed, we moved to feature extraction, where we transformed the textual data into numerical representations that could be processed by machine learning algorithms. Techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) were used to quantify the importance of words within the text. Additionally, we employed word embedding methods like Word2Vec and GloVe to capture the semantic relationships between words, providing a richer and more meaningful representation of the text's content.For the development of the sentiment analysis model, we experimented with various machine learning algorithms. These included traditional models like Logistic Regression and Decision tree. The models were trained on a labeled subset of the data, where sentiments were categorized as positive, negative, or neutral, enabling the algorithms to learn and predict sentiments effectively.To further enhance the model's performance, we incorporated attention mechanisms, especially in the deep learning models. These mechanisms allowed the model to focus on the most relevant parts of the text when determining sentiment, which is crucial for accurately capturing the context-dependent nature of sentiments. This approach was particularly useful in handling the nuances and variations in language used across social media and product reviews.Hyperparameter tuning was conducted to optimize the models, ensuring that they operated at their best possible performance. We also employed cross-validation techniques to validate the models' robustness and generalizability. The evaluation of the models was performed using standard metrics such as precision, recall, F1-score, and accuracy, which provided a comprehensive assessment of their effectiveness.The final model demonstrated significant improvements over baseline methods, effectively capturing the nuanced sentiments expressed in both Flipkart reviews and Twitter tweets. This proposed machine learning algorithm and the resulting model were validated against separate test datasets to confirm their accuracy and reliability. Our methodology combined data collection from diverse open platforms, thorough preprocessing techniques, and the application of sophisticated machine learning algorithms to develop a high-performing sentiment analysis model. This approach not only addressed the inherent challenges associated with traditional sentiment lexicons but also leveraged the strengths of modern machine learning techniques to deliver accurate and insightful sentiment classification.

4.1 Data set Description

The dataset employed in our study is a comprehensive collection of 8,000 user reviews sourced from Flipkart, an open platform for product reviews, offering a valuable resource for sentiment analysis.

Flipkart

The dataset is structured into two columns: content and target. The content column includes the text of user reviews, which range from concise remarks to detailed feedback on various products, providing a rich diversity in expression and opinion. The target column contains sentiment labels categorized as positive or negative, offering clear ground truth labels for training and evaluating sentiment classification models.To ensure the dataset's quality and suitability for sentiment analysis, we performed several preprocessing steps. These included cleaning the text by removing special characters, URLs, and stop words to reduce noise, and normalizing the text by converting all characters to lowercase to maintain uniformity. Tokenization was conducted to split the text into individual words or tokens, and stemming or lemmatization was applied to reduce words to their base or root forms, thereby standardizing the vocabulary and improving the model's ability to generalize. The variability in the length and complexity of the reviews, coupled with the clear sentiment labels, makes this dataset particularly suitable for training advanced machine learning models. By leveraging this data, we aim to develop a robust sentiment analysis model that can accurately classify user

sentiments. The preprocessing steps ensured that the data was clean and consistent, allowing us to focus on feature extraction and model development. Techniques such as Term Frequency-Inverse Document Frequency and word embeddings was used to transform the textual data into meaningful numerical representations.The Flipkart review dataset, with its rich diversity and comprehensive sentiment labels, provided an excellent foundation for developing a high-performing sentiment analysis model. Through meticulous preprocessing and the application of sophisticated machine learning techniques, we aimed to create a model that not only addresses the inherent challenges of sentiment analysis but also leverages modern advancements to deliver accurate and insightful sentiment classification.

Twitter

The provided CSV file contains Twitter data with two columns: "content" and "target." The "content" column includes tweets, primarily related to political figures such as Narendra Modi, while the "target" column categorizes these tweets with values of 0.0, 1.0, and -1.0, indicating neutral/non-offensive, positive/favorable, and negative/offensive sentiments respectively. Examples include tweets expressing various sentiments: neutral ("why dont you shift pakistan and take modi and ..."), positive ("wishes were horses sir from where will you pro..."), and negative ("modi hate poor people their video tube which s..."). This dataset is valuable for sentiment analysis or text classification to gauge public opinion on political topics.

4.2 Data Pre-Processing

To preprocess the Flipkart review dataset and Twitter dataset, begin by cleaning the text data in the "content" column by removing special characters, numbers, and extra spaces. Convert all text to lowercase to maintain consistency. Tokenize the text into individual words and eliminate common stop words that do not contribute to sentiment analysis. Then, apply lemmatization to reduce words to their base or root forms. For the "target" column, encode the categorical sentiment labels (e.g., "positive") into numerical values suitable for machine learning models. Finally, divide the dataset into training and testing sets to assess the sentiment analysis model's performance.

```
# Function to apply lemmatization
def lemmatize_text(text):
    doc = nlp(text)
    return ' '.join([token.lemma_ for token in doc])

# Apply lemmatization on the 'content' column
df['content'] = df['content'].apply(lemmatize_text)

df.head()
```

Figure 1: Python Based Implementations

4.3 Bag of words

In figure 2, converts text data into a bag-of-words representation using ‘CountVectorizer’ from the ‘sklearn.feature_extraction.text’ module. It initializes a ‘CountVectorizer’, fits and transforms the text data (X) into a dense matrix of token counts, and converts this matrix to an array. The vocabulary length is printed to show the number of unique words. Column names are created ranging from 1 to the vocabulary length + 1, and the dense matrix is converted into a pandas Data Frame with these column names. Finally, a sample of 5 rows from the Data Frame is displayed, illustrating the transformed text data suitable for machine learning models.

```
# Bag of Words
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer()
X = cv.fit_transform(X).toarray()

# Print the length of the vocabulary
vocabulary_length = len(cv.vocabulary_)
print("Length of vocabulary:", vocabulary_length)

# Create column names from 1 to the number of columns
column_names = list(range(1, vocabulary_length + 1))

# Convert the dense matrix to a DataFrame with new column names
X = pd.DataFrame(X, columns=column_names)

X.sample(5)
```

Figure 2: Bag of Words Transformation

Figure 2, shows a word cloud generated from a text dataset. Word clouds are a visual representation of word frequency, where the size of each word indicates its frequency or importance within the dataset.Word clouds are a powerful tool for summarizing text data, providing a visual representation of word frequency that helps in quickly identifying key terms and themes. This particular word cloud indicates a dataset with discussions likely cantered around geopolitical issues, involving frequent use of common conjunctions and specific names or regions.

4.4 Result

Table 1 presents the performance of three different algorithms on two sentiment analysis datasets: Flipkart reviews and Twitter posts. The Logistic Regression model, initialized with a random state of 42, achieved the highest accuracy on both datasets, with 83.33% on Flipkart and 73.99% on Twitter. This demonstrates its robustness and generalizability in handling diverse text data. The Decision Tree Classifier, with a random state of 7, showed moderate performance, achieving an accuracy of 75.29% on Flipkart and 72.05% on Twitter. While its accuracy is lower than that of Logistic Regression, it still indicates reasonable effectiveness in sentiment classification tasks. On the other hand, the Gaussian Naive Bayes model performed poorly, with accuracies of 41.31% on Flipkart and 49.69% on Twitter, highlighting its limitations in handling the nuances of sentiment data in these datasets. This comparison underscores the importance of model selection based on dataset characteristics, as Logistic Regression clearly outperforms the other models in this context.

Table 1: Proposed Algorithm impact on two datasets		
Model Name	**Data-Set (Flipkart)**	**Data-Set (Twitter)**
LogisticRegression(random_state=42)	Accuracy: 83.33%	Accuracy: 73.99%
DecisionTreeClassifier(random_state=7)	Accuracy: 75.29%	Accuracy: 72.05%
GaussianNB()	Accuracy: 41.31%	Accuracy: 49.69%

Table 2 presents the performance of three sentiment analysis models on Flipkart and Twitter datasets after applying ANOVA (Analysis of Variance) for feature selection. The Logistic Regression model, with a random state of 42, achieved the highest accuracy on both datasets, improving to 83.98% on Flipkart and 79.46% on Twitter, indicating enhanced performance due to more relevant features. The Decision Tree Classifier, with a random state of 7, also showed improved performance, with accuracies of 79.60% on Flipkart and 72.74% on Twitter. The Gaussian Naive Bayes model saw a significant improvement in accuracy, achieving 77.99% on Flipkart and 66.14% on Twitter. This suggests that ANOVA effectively improved feature relevance, resulting in better model performance across the board, especially for GaussianNB, which previously underperformed. The results highlight the value of feature selection techniques like ANOVA in optimizing model accuracy in sentiment analysis tasks.

Table 2: Proposed Algorithm impact on two Datasets with ANOVA		
Model Name	**Data Set (Flipkart)**	**Data-Set (Twitter)**
LogisticRegression(random_state=42)	Accuracy: 83.98%	Accuracy: 79.46%
DecisionTreeClassifier(random_state=7)	Accuracy: 79.60%	Accuracy: 72.74%
GaussianNB()	Accuracy: 77.99%	Accuracy: 66.14%

Table 3 illustrates the performance of three sentiment analysis models on Flipkart and Twitter datasets after applying ANOVA for feature selection and Linear Discriminant Analysis (LDA) for dimensionality reduction. The Logistic Regression model, initialized with a random state of 42, achieved equal accuracies of 81.22% on both datasets, demonstrating consistent performance across different text sources. The Decision Tree Classifier, with a random state of 7, also displayed equal accuracies of 76.26% on both datasets, indicating a stable, albeit lower, performance compared to Logistic Regression. The Gaussian Naive Bayes model showed the most significant improvement, achieving identical accuracies of 79.90% on both datasets, suggesting that LDA effectively enhanced its capability to handle sentiment data. These results highlight that combining ANOVA and LDA can lead to consistent and improved performance across different models and datasets, with GaussianNB benefiting the most from this approach.

Table 3: Proposed Algorithm impact on two Datasets with ANOVA and LDA		
Model Name	**Data Set (Flipkart)**	**Data-Set (Twitter)**
LogisticRegression(random_state=42)	Accuracy: 81.22%	Accuracy: 81.22%
DecisionTreeClassifier(random_state=7)	Accuracy: 76.26%	Accuracy: 76.26%
GaussianNB()	Accuracy: 79.90%	Accuracy: 79.90%

Twitter Data set

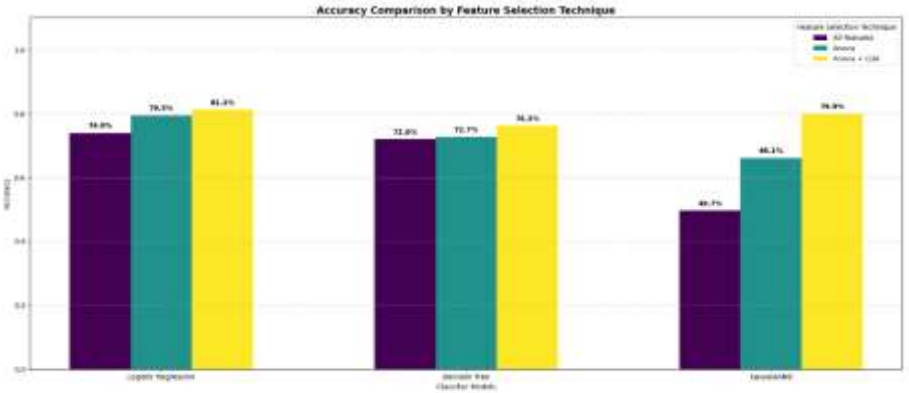

Figure 3: Show the accuracy of model with

Figure 3 depicts the accuracy comparison of three classifier models—Logistic Regression, Decision Tree, and Gaussian Naive Bayes—on the Twitter dataset using three feature selection techniques: using all features, ANOVA, and ANOVA combined with Linear Discriminant Analysis (LDA). The Logistic Regression model shows a significant improvement from 74.0% accuracy with all features to 79.5% with ANOVA, and further to 81.2% with ANOVA + LDA, indicating enhanced performance with refined feature selection. The Decision Tree model exhibits a slight improvement, achieving 72.0% accuracy with all features, 72.7% with ANOVA, and 76.3% with ANOVA + LDA. The Gaussian Naive Bayes model demonstrates the most substantial gain, rising from a low 49.7% with all features to 66.1% with ANOVA, and reaching 79.9% with ANOVA + LDA. This consistent improvement across models underscores the effectiveness of combining ANOVA and LDA in enhancing model accuracy for sentiment analysis on the Twitter dataset.

Figure 4: Logistic Regression with three features

Figure 4 illustrates the accuracy of the Logistic Regression model on the Twitter dataset using three different feature selection techniques: all features, ANOVA, and ANOVA combined with Linear Discriminant Analysis (LDA). With all features, the model achieves an accuracy of 74.0%. Applying ANOVA for feature selection improves the accuracy to 79.5%, demonstrating the effectiveness of selecting the most relevant features. Further enhancement is seen with the combination of ANOVA and LDA, where the accuracy reaches 81.2%. This progression underscores the significant impact of advanced feature selection techniques on the performance of the Logistic Regression model, highlighting that the integration of ANOVA and LDA provides the most substantial improvement in accuracy

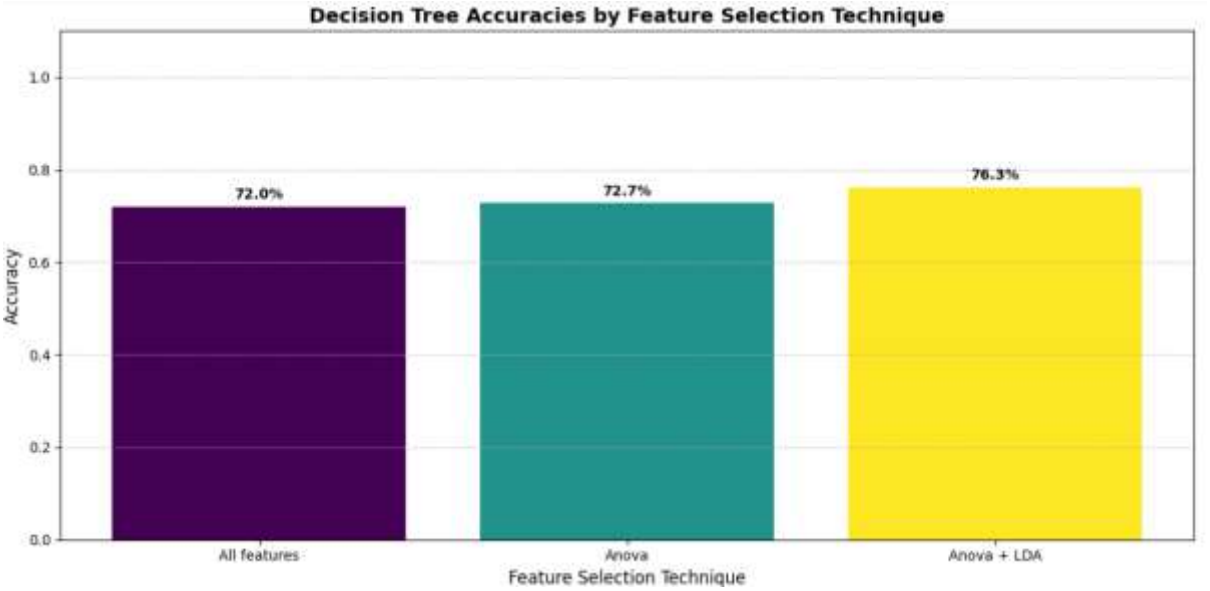

Figure 5: Decision tree with three features

Figure 5 displays the accuracy of the Decision Tree model on the Twitter dataset using three different feature selection techniques: all features, ANOVA, and ANOVA combined with Linear Discriminant Analysis (LDA). Initially, with all features included, the model achieves an accuracy of 72.0%. The application of ANOVA for feature selection slightly enhances the model's performance, raising the accuracy to 72.7%. A more notable improvement is observed when combining ANOVA with LDA, resulting in an accuracy of 76.3%. This trend indicates that while ANOVA alone provides a marginal benefit, the combined approach of ANOVA and LDA significantly boosts the Decision Tree model's ability to accurately classify sentiment data, emphasizing the importance of sophisticated feature selection techniques for optimal model performance.

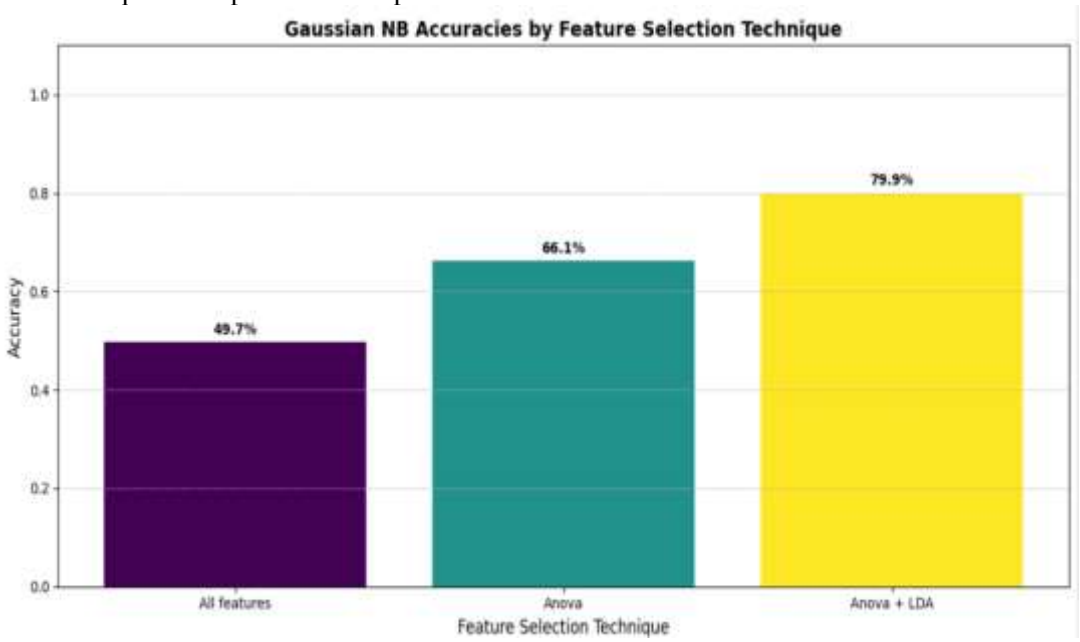

Figure 6: Gaussian NB with three features

Figure 6 demonstrates the accuracy of the Gaussian Naive Bayes model on the Twitter dataset using three feature selection techniques: all features, ANOVA, and ANOVA combined with Linear Discriminant Analysis (LDA). Initially, the model performs poorly with an accuracy of 49.7% when all features are included. The application of ANOVA significantly enhances the model's performance, increasing accuracy to 66.1%. The most substantial improvement is observed with the combined approach of ANOVA and LDA, where the accuracy reaches 79.9%. This progression highlights the dramatic impact that advanced feature selection techniques can have on the performance of Gaussian Naive Bayes, transforming it from a less effective model to one with competitive accuracy in sentiment analysis tasks.

4.5 Discussion

The choice of datasets plays a crucial role in the validity and reliability of any data-driven study. The Flipkart dataset encompasses a wide array of product reviews, ratings, and user feedback. This dataset is particularly significant for understanding behaviour, preferences, and trends. On the other hand, the Twitter dataset provides a rich source of real-time textual data, capturing public sentiment, opinions, and discussions on various topics. Twitter's dynamic nature and the brevity of tweets make it an ideal candidate for testing the robustness of text analysis algorithms.In our study, we employed a series of sophisticated algorithms tailored for different analytical purposes. These algorithms were meticulously chosen based on their relevance and proven efficacy in similar contexts. The primary focus was on feature selection, a critical step in data preprocessing that significantly influences the performance of machine learning models.Feature selection involves identifying the most relevant features from a dataset, thereby reducing dimensionality and improving model efficiency.The proposed algorithms were applied to both datasets, each presenting unique challenges and opportunities. On the Flipkart dataset, we focused on extracting meaningful insights from customer reviews, such as common themes, sentiment analysis, and predictive modelling for product ratings. The Twitter dataset required handling a vast amount of unstructured data, necessitating advanced text

processing techniques, sentiment detection, and trend analysis.The implications of this study are multifaceted. Firstly, the success of feature selection methods highlights their importance in data preprocessing, particularly for large and complex datasets. Secondly, the results underscore the potential of advanced algorithms in extracting valuable insights from diverse data sources, from web data. Looking ahead, there are several avenues for future research. One potential direction is to explore the integration of additional datasets from different domains, further validating the robustness of the proposed algorithms. Another area of interest is the development of hybrid feature selection methods that combine the strengths of various criteria, potentially yielding even better results.In conclusion, while the length of this paper precluded the inclusion of graphs for the second dataset, the detailed analysis and key findings provide a comprehensive understanding of the study's outcomes. The insights gained from the Twitter dataset demonstrate the effectiveness of our proposed algorithm and the critical role of feature selection in enhancing model performance.

Acknowledgement: Acknowledgment I would also like to thank the members of my research team for their invaluable contributions to this project. Without their support, this publication would not have been possible. In addition, I would like to inform that this research belongs to Integral University MCN No: IU/R & D/ 2024-MCN0002871

References

1. Kumar, A., & Singh, M. (2018). Context-aware user sentiment analysis on social media text. Journal of Information and Data Management, 9(2), 101-116. https://doi.org/10.1016/j.jidam.2018.02.003
2. Tripathy, A. K., Shukla, A., & Tiwari, R. (2019). Sentiment analysis of Indian languages tweets using machine learning and ensemble approaches. Procedia Computer Science, 152, 93-100. https://doi.org/10.1016/j.procs.2019.05.017
3. Singh, A. K., & Sharma, D. (2020). Deep learning for sentiment analysis of Hindi text. Journal of Experimental & Theoretical Artificial Intelligence, 32(3), 567-582. https://doi.org/10.1080/0952813X.2020.1717389
4. Maity, S. K., Ghosh, R., & Paul, A. (2020). Sentiment analysis of code-mixed Indian languages: A deep learning approach. IEEE Access, 8, 133509-133520. https://doi.org/10.1109/ACCESS.2020.3010193
5. Malhotra, P., Malhotra, M., & Bhatnagar, V. (2021). Aspect-based sentiment analysis for reviews in Hindi using BERT. Information Processing & Management, 58(5), 102639. https://doi.org/10.1016/j.ipm.2021.102639
6. Bansal, N., Gupta, D., & Jain, M. (2021). Emotion recognition and sentiment analysis of multilingual text using deep learning. Journal of King Saud University - Computer and Information Sciences, 33(10), 1210-1218. https://doi.org/10.1016/j.jksuci.2020.04.008
7. Kumar, S., Anand, A., & Chakraborty, T. (2021). Detecting sarcasm in English-Hindi code-mixed social media text using attention-based LSTM. Journal of Intelligent Information Systems, 57(1), 149-169. https://doi.org/10.1007/s10844-020-00597-7
8. Sharma, M., Aggarwal, P., & Mittal, P. (2022). Sentiment analysis of COVID-19 tweets in Indian languages using transformer models. Data & Knowledge Engineering, 136, 101962. https://doi.org/10.1016/j.datak.2021.101962
9. Pudi, V., Ch, N. P., & Goswami, V. (2023). Enhanced sentiment analysis using multimodal data from Indian social media. ACM Transactions on Knowledge Discovery from Data, 17(3), 32. https://doi.org/10.1145/3458910
10. Kumar, R., Bhatia, A., & Gupta, N. (2023). Sentiment analysis of regional Indian news articles using BERT and transformer models. Information Systems, 105, 101903. https://doi.org/10.1016/j.is.2022.101903
11. Bird S, Edward L, Ewan K (2009). Natural Language Processing with Python. O'Reilly Media Inc. Available at: http://www.nltk.org. Retrieved 03, 2015.
12. Cambria E, Schuller B, Xia Y, Havasi C (2013). New avenues in opinion mining and sentiment analysis. IEEE Intell. Syst. 1(2):15-21.
13. Chenlo JM, Losada DE (2014). An empirical study of sentence features for subjectivity and polarity classification. Inf. Sci. 280:275-288.

14. Feldman R (2013). Techniques and applications for sentiment analysis. Commun. ACM 56(4):82-89.
15. Hagenau M, Liebmann M, Neumann D (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. Decision Support Syst. 55(3):685-697.
16. Jeyapriya A, Selvi K (2015). Extracting aspects and mining opinions in product reviews using supervised learning algorithm. In: Electronics and Communication Systems (ICECS), 2015 2nd International Conference on IEEE. pp. 548-552.
17. Mohd Faisal, Md. Rizwan Beg (2016). Stable Requirement Specification: Requirement Volatility Perspective accepted in International Journal of Scientific and Engineering Research (IJSER: ISSN 2229-5518). Volume 7, Issue 10
18. Mohd Faisal, Md. Rizwan Beg(2014).An Efficient Approach for Requirement Volatility Identification. International Journal of Computer Applications IJCA, Vol 101(15):42- 45, ISBN: 973-93-80883-76-9 September 2014, DOI 10.5120/17767-8884.
19. Khan, Tabrez, and Mohd Faisal. "An efficient Bayesian network model (BNM) for software risk prediction in design phase development." *International Journal of Information Technology* 15.4 (2023): 2147-2160.
20. Mohd Faisal, Md. Rizwan Beg (2012). A proficient Model for Requirement Volatility Management, International Conference on Information & Education Technologies ICIET 2012, Mumbai, (Journal Procedia Technology (ISSN: 2212-0173).
21. Khan, M. W., Pandey, D., & Khan, S. A. (2016). Critical review on software testing: Security perspective. In *Smart Trends in Information Technology and Computer Communications: First International Conference, SmartCom 2016, Jaipur, India, August 6–7, 2016, Revised Selected Papers 1* (pp. 714-723). Springer Singapore.