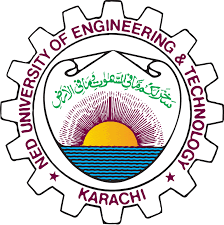
**PROJECT REPORT**

**DEPARTMENT: COMPUTER SCIENCE AND INFORMATION TECHNOLOGY**

**DATA MINING (CT-377)**

**Instructor Name: Sir Rohail Qamar**

**Topic:** **Sentiment Analysis for Product Review**

**GROUP MEMBERS**

SYED MUHAMMAD ABDULLAH

SYED FARHAN ALI RAZA

UZAIR AHMED

QURAT-UL-AIN AKHTER

HAJRA RAO

**TABLE OF CONTENTS**

[Introduction 3](#_Toc182697593)

[1.2 Abstract 3](#_Toc182697595)

[1.2 Aims And Objectives 3](#_Toc182697595)

[1.3 Methodology 3](#_Toc182697596)

Literature Review 5

[2.1 Introduction 4](#_Toc182697597)

[2.2 Importance and Role of Sentiment Analysis in Business Decision-Making 4](#_Toc182697598)

[2.3 Methodologies in Sentiment Analysis 4](#_Toc182697599)

[2.3.2 Machine Learning Algorithms in Sentiment Analysis 4](#_Toc182697600)

[2.3.2.1 Naive Bayes 4](#_Toc182697601)

[2.3.2.2 Support Vector Machines 5](#_Toc182697602)

[2.3.2.3 Maximum Entropy 5](#_Toc182697603)

[2.4 Challenges in Sentiment Analysis 5](#_Toc182697604)

[2.5 The Impact of the World Wide Web on Sentiment Analysis 5](#_Toc182697605)

[2.6 Advances and Innovations in Sentiment Analysis Techniques 5](#_Toc182697606)

[2.7 Evolution of Research in Sentiment Analysis 6](#_Toc182697607)

[2.8 Keyword Co-occurrence Analysis in Sentiment Research 6](#_Toc182697608)

[2.9 Future Directions and Areas for Improvement 6](#_Toc182697609)

[2.10 Conclusion 7](#_Toc182697610)

Results and Discussions 9

3.1 Dataset Preprocessing and overview 9

3.2 Model Performance 9

3.2.1 XGBoost 9

3.2.2 Logistic Regression 9

3.2.3 Multinomial Naïve Bayes 10

3.2.4 Bernoulli Naïve Bayes 10

3.2.5 Support vactor Machine 10

3.3 Key Observations 10

3.4 Comparative Visualization 10

Conclusion And Future Work 11

4.1 Conclusion 11

4.1 Future Work 11

References……... 12

**CHAPTER 1**

**INTRODUCTION**

### ABSTRACT

This project focuses on performing sentiment analysis on product reviews, aiming to classify customer feedback as positive or negative. By analyzing user-generated content, the system helps businesses gain insights into customer satisfaction and product performance. The project involves data preprocessing, text processing, feature extraction using TF-IDF, and training multiple machine learning models: Logistic Regression, Support Vector Machine (SVM), XGBoost, and Naive Bayes. The evaluation of these models is performed using metrics such as accuracy and classification reports, providing a detailed understanding of each model’s performance. This analysis can assist companies in better understanding consumer needs and improving their products accordingly.

### AIMS AND OBJECTIVES

* To develop a sentiment analysis model for classifying product reviews as positive or negative.
* To preprocess raw text data by removing noise, stop words, and applying lemmatization.
* To implement feature extraction using TF-IDF for converting text data into numerical format.
* To train and compare the performance of four machine learning models: Logistic Regression, SVM, XGBoost, and Naive Bayes.
* To evaluate the performance of the models using metrics such as accuracy and classification reports.

### METHODOLOGY

* **Data Collection:** The dataset consists of product reviews gathered from an online source.
* **Data Preprocessing:** The text data is cleaned by removing special characters, stop words, and applying lemmatization.
* **Feature Extraction:** TF-IDF vectorization is used to convert the preprocessed text into numerical features suitable for machine learning.
* **Model Training:** Four different machine learning models (Logistic Regression, SVM, XGBoost, and Naive Bayes) are trained on the feature-extracted data.
* **Evaluation:** The models’ performances are assessed using metrics such as accuracy and classification reports to identify the most effective classifier.
* **Deployment:** The final selected model can be used for real-time sentiment analysis of new product reviews. We used streamlit for developing the application for deployment.

**CHAPTER 2**

**LITERATURE REVIEW**

### Introduction

In the digital age, the proliferation of online content has transformed the way individuals and organizations communicate, share opinions, and make decisions. Sentiment analysis, also known as opinion mining, has emerged as a critical field within natural language processing (NLP) that focuses on extracting and analyzing subjective information from textual data. By leveraging advanced computational techniques, sentiment analysis enables the identification and categorization of sentiments expressed in various forms of communication, such as social media posts, product reviews, and customer feedback. The three papers collectively explore the importance, challenges, and evolving methodologies of sentiment analysis in natural language processing, emphasizing its role in data analytics and the need for improved techniques to handle user-generated content.

### Importance and Role of Sentiment Analysis in Business Decision-Making

In the first paper [1], researchers have noted that data analytics is widely utilized across various industries and organizations to enhance business decision-making. They emphasize that by applying analytics to both structured and unstructured data, enterprises can significantly transform their planning and decision-making processes. In the authors' view, sentiment analysis, also referred to as opinion mining, plays a crucial role in everyday decision-making. They argue that these decisions can range from purchasing products, such as mobile phones, to reviewing movies and making investment choices, all of which can have a substantial impact on daily life.

### Methodologies in Sentiment Analysis

* + 1. **Overview of Common Techniques**

The authors present a detailed survey of various methodologies and approaches to sentiment analysis, aiming to enhance understanding of the available techniques.

### Machine Learning Algorithms in Sentiment Analysis

Researchers highlight that common machine learning algorithms, such as Naïve Bayes, Support Vector Machines, and Maximum Entropy, typically provide limited classification categories ranging between positive and negative.

### Naive Bayes

Naive Bayes is one of the common algorithms discussed in the context of sentiment analysis

### Support Vector Machines

Support Vector Machines are also highlighted as a common approach in sentiment classification.

### Maximum Entropy

Maximum Entropy is mentioned as another algorithm used for sentiment analysis.

### Challenges in Sentiment Analysis

* + 1. **Polarity Shift**

The authors observe that sentiment analysis encounters several challenges, including polarity shift.

* + 1. **Accuracy Issues**

They point out accuracy-related issues that hinder effective sentiment analysis.

* + 1. **Binary Classification Problems**

The binary classification problem is noted as a significant challenge in sentiment analysis.

* + 1. **Data Sparsity**

Data sparsity is another challenge that the authors highlight as affecting sentiment analysis.

### The Impact of the World Wide Web on Sentiment Analysis

* + 1. **Data Generation through Social Media and E-commerce**

In the second paper [2], researchers define sentiment analysis as the process of mining data, views, reviews, or sentences to predict the emotion conveyed in the text through natural language processing (NLP). They observe that in recent years, the World Wide Web (WWW) has emerged as a vast source of raw data generated by users.

* + 1. **Value of User-Generated Content**

Researchers note that through social media platforms and e-commerce websites, such as Facebook, Twitter, Amazon, and Flipkart, users conveniently share their views and feelings. The authors argue that this growing volume of raw data is an extremely valuable source of information for decision-making processes.

### Advances and Innovations in Sentiment Analysis Techniques

* + 1. **Recent Developments**

While sentiment analysis is fundamentally text-based, the authors acknowledge that there are significant challenges in accurately determining the polarity of sentences. They contend that there is a pressing need to develop better solutions that yield improved results compared to previous approaches or techniques used for polarity detection.

* + 1. **Proposed New Techniques**

In their paper [2], the researchers present a detailed survey of various techniques and approaches employed in sentiment analysis, along with a new technique proposed by the authors.

### Evolution of Research in Sentiment Analysis

* + 1. **Growth of Academic Interest**

In the third paper [3], researchers have identified sentiment analysis as one of the prominent research hotspots in the field of natural language processing, garnering significant attention from the academic community.

* + 1. **Historical Trends and Emerging Hotspots**

The authors highlight that the volume of research papers published in this area is steadily increasing, and they observe that there have been few surveys works that leverage keyword co- occurrence analysis in the context of sentiment analysis.

### Keyword Co-occurrence Analysis in Sentiment Research

* + 1. **Methodology and Applications**

Consequently, the researchers present a paper [3] that focuses on surveying sentiment analysis with an emphasis on the evolution of research methods and topics. They incorporate keyword co-occurrence analysis alongside a community detection algorithm.

* + 1. **Insights from Community Detection Algorithms**

The authors assert that this survey not only compares and analyzes the connections between research methods and topics over the past two decades but also uncovers emerging hotspots and trends over time.

### Future Directions and Areas for Improvement

* + 1. **Identifying Limitations**

Furthermore, they claim that this paper [3] offers broad practical insights into the methods and topics of sentiment analysis while also identifying technical directions and limitations.

* + 1. **Recommendations for Future Research**

Future areas of work in research include ways to handle change in sentiment, improve the accuracy of sentiment classification, and help overcome the problem of limited data. The further integration of various data types, such as text, images, and audio, will also be needed in the improvement of sentiment analysis. Cultural and language differences will be an important consideration in making better and more accurate models for global sentiment analysis.

### Conclusion

Sentiment analysis is used to extract from text people's opinions and feelings about products, services, concepts, ideas, and so on, which could come in the form of posts in social media, reviews in blogs, or comments about some events. Techniques such as Naïve Bayes, SVM, and Maximum Entropy classify sentiment as well as possible, yet problems such as changes in the meaning of the sentiment, accuracy problems, and a lack of data persist. The sheer quantity of user-generated content online is an opportunity but also a challenge. New methodologies, such as keyword co-occurrence analysis and community detection, may hold promise, but much remains to be done in fully understanding sentiment in text.

**CHAPTER 3**

**RESULTS AND DISCUSSIONS**

The sentiment analysis project utilized various machine learning models to classify customer reviews into three classes: positive, neutral, and negative sentiments. Key steps included preprocessing the text, balancing the dataset using SMOTE, and evaluating the performance of models based on accuracy, precision, recall, and F1-score. The following sections discuss the findings and comparative performance of the models:

* 1. **DATASET OVERVIEW AND PREPROCESSING**

The dataset initially had missing values, which were handled by dropping incomplete rows. Reviews were cleaned by removing HTML tags, punctuation, and stop words, followed by lemmatization. The processed dataset revealed an imbalance in class distribution, with a significantly higher number of positive reviews. This imbalance was addressed through downsampling for positive reviews and upsampling using SMOTE.

The reviews were transformed into numerical features using TF**-**IDFvectorization, selecting the top 7500 features based on frequency and importance.

* 1. **MODEL PERFORMANCE**

Multiple machine learning algorithms were tested for sentiment classification, including XGBoost, LogisticRegression, MultinomialNaiveBayes, BernoulliNaiveBayes, and SupportVectorMachine **(**SVM**)**. The results are summarized below:

1. **XGBoost**

* Accuracy: 76%
* Precision: 76%
* Recall: 76%
* F1-Score: 76%
* Training Time: 258 minutes

XGBoost performed consistently across metrics, demonstrating its ability to capture complex patterns. However, it had the highest computational cost.

1. **Logistic Regression**

* Accuracy: 79%
* Precision: 80%
* Recall: 79%
* F1-Score: 79%
* Training Time: 25 minutes

Logistic Regression emerged as the best-performing model, offering the highest accuracy and F1-score. Its training time was also significantly lower than that of XGBoost and SVM.

1. **Multinomial Naïve Bayes**

* Accuracy: 73%
* Precision: 76%
* Recall: 73%
* F1-Score: 74%
* Training Time: 4 minutes

Multinomial Naive Bayes provided competitive results for precision and F1-score, but its overall accuracy lagged behind Logistic Regression and XGBoost. It remains a computationally efficient option.

1. **Bernoulli Naïve Bayes**

* Accuracy: 65%
* Precision: 67%
* Recall: 65%
* F1-Score: 65%
* Training Time: 5 minutes

Bernoulli Naive Bayes was the least effective, likely due to the nature of text data and binary feature representation.

1. **Support Vector Machine (SVM)**

* Accuracy: 79%
* Precision: 80%
* Recall: 79%
* F1-Score: 79%
* Training Time: 180 minutes

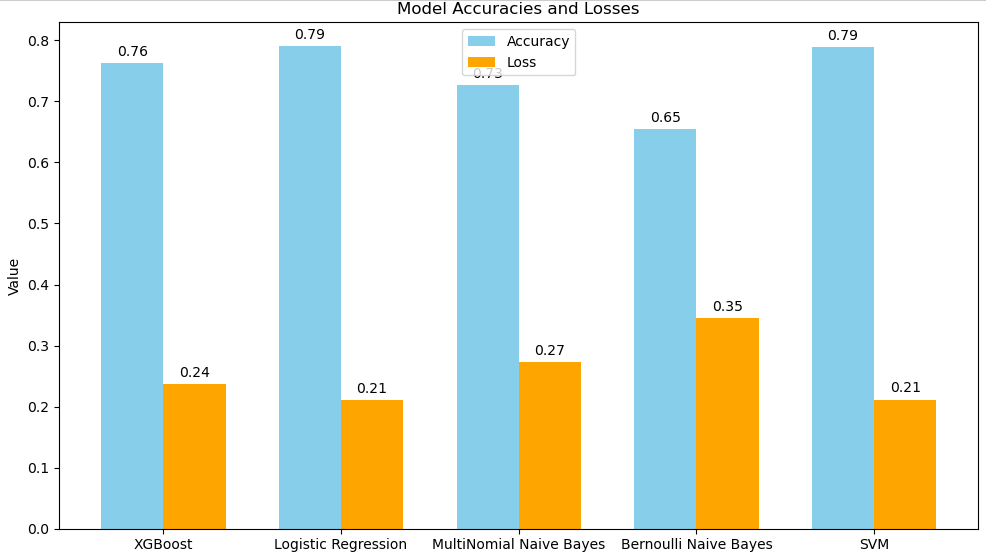
SVM matched Logistic Regression in terms of accuracy and other metrics but incurred significantly higher computational costs.

1. **KEY OBSERVATIONS**

* Logistic Regression achieved the best trade-off between performance and computational efficiency, making it the most suitable model for real-time applications.
* XGBoost and SVM also delivered high accuracy but required longer training times, which might limit their scalability.
* Both Naive Bayes models exhibited moderate performance, with Multinomial Naive Bayes outperforming its Bernoulli counterpart.
* The SMOTE technique successfully balanced the training dataset, improving the model's ability to predict minority classes.
* The visualization of a Word Cloud highlighted key terms contributing to the sentiment classification, offering interpretability for insights into review content.

1. **COMPARATIVE VISUALIZATION**

A bar chart was used to compare model accuracy and loss, while a tabular format highlighted training times and detailed metrics (accuracy, precision, recall, and F1-score). The findings emphasize the trade-offs between accuracy and computational complexity.



**CHAPTER 4**

**CONCLUSION AND FUTURE WORK**

* 1. **CONCLUSION**

The sentiment analysis project compared various machine learning models to classify customer reviews into positive, neutral, and negative sentiments. Among the tested models, Logistic Regression and Support Vector Machine (SVM) achieved the highest performance across all metrics, including accuracy, precision, recall, and F1-score, indicating a tie in terms of classification effectiveness. However, the training time for Logistic Regression was significantly lower at 25 minutes, compared to SVM's 180 minutes.

Given the trade-off between performance and computational efficiency, Logistic Regression was selected as the final model. Its faster training time makes it more suitable for real-time applications and scalable implementations without compromising classification accuracy.

* 1. **Future Work**

Looking ahead, there are several key areas for further research and development in sentiment analysis. First, the challenge of polarity shifts needs to be addressed more effectively, as current techniques struggle to capture context-dependent sentiment. Second, improving the accuracy of sentiment classification, particularly in the face of data sparsity and ambiguous expressions, remains a critical task. Advanced approaches, such as deep learning and neural networks, could provide solutions to these issues. Moreover, the integration of multimodal data (e.g., combining text with images, audio, or video) holds potential for richer sentiment analysis, especially in the context of social media. Additionally, further exploration into the application of keyword co-occurrence analysis and community detection algorithms could reveal emerging trends and improve the detection of nuanced sentiments. Lastly, a more comprehensive understanding of cultural and linguistic differences is essential for enhancing sentiment analysis models in a globalized, multi-lingual digital landscape. Addressing these challenges will pave the way for more accurate, context-aware, and scalable sentiment analysis tools in the future.

### References

[1] A. M. Abirami and V. Gayathri, "A survey on sentiment analysis methods and approach," 2016 Eighth International Conference on Advanced Computing (ICoAC), Chennai, India, 2017, pp. 72-76, doi: 10.1109/ICoAC.2017.7951748.

keywords: {Sentiment analysis;Ontologies;Feature extraction; Dictionaries; Blogs; Data mining; Data analysis; Data Analytics; sentiment Analysis; Decision making}

[2] N. Sultana, P. Kumar, M. R. Patra, S. Chandra, and S. K. S. Alam, “Sentiment Analysis for Product Review,” ICTACT Journal on Soft Computing, vol. 9, no. 3, pp. 1913-1919, Apr. 2019, doi: 10.21917/ijsc.2019.0266.

[3] Cui, J., Wang, Z., Ho, SB. et al. “Survey on sentiment analysis: evolution of research methods and topics.” Artif Intell Rev56, 8469–8510 (2023). https://doi.org/10.1007/s10462-022-10386-z