A

Mini Project Report on

## Sentiment Analysis Using NLP

Submitted in partial fulfillment of the requirements for the degree of

**BACHELOR OF ENGINEERING**

IN

### Computer Science & Engineering

### (Artificial Intelligence & Machine Learning)

by

Maitreyi Phadke (22106007)

Gauri Ramekar (22106068)

Arya Raut (22106075)

Chinmay Sawant (22106017)

Under the guidance of

## Prof. Mahesh Pawaskar



### Department of Computer Science & Engineering

### (Artificial Intelligence & Machine Learning)

**A. P. Shah Institute of Technology**

**G. B. Road, Kasarvadavali, Thane (W)-400615**

**University Of Mumbai**

**2024-2025**

## A. P. SHAH INSTITUTE OF TECHNOLOGY

## CERTIFICATE

This is to certify that the project entitled “**SENTIMENT ANALYSIS”** is a bonafide work of Maitreyi Phadke (22106007), Gauri Ramekar (22106068), Arya Raut (22106075), Chinmay Sawant (22106017) submitted to the University of Mumbai in partial fulfillment of the requirement for the award of **Bachelor of Engineering** in **Computer Science & Engineering (Artificial Intelligence & Machine Learning).**

|  |  |
| --- | --- |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| Prof. Mahesh Pawaskar | Dr. Jaya Gupta |
| Mini Project Guide | Head of Department |

## 

## A. P. SHAH INSTITUTE OF TECHNOLOGY

## Project Report Approval

This Mini project report entitled “**Sentiment Analysis*”*** by **Maitreyi Phadke, Gauri Ramekar, Arya Raut, Chinmay Sawant**is approved for the degree of ***Bachelor of Engineering*** in ***Computer Science &Engineering***, (AIML) ***2024-25***.

##### External Examiner:

##### Internal Examiner:

Place: APSIT, Thane

Date:

**Declaration**

##### We declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

|  |  |  |  |
| --- | --- | --- | --- |
| **Maitreyi Phadke** | **Gauri Ramekar** | **Arya Raut** | **Chinmay Sawant** |
| (22106007) | (22106068) | (22106075) | (221060017) |

#### ABSTRACT

Sentiment analysis, or opinion mining, is a pivotal subfield of Natural Language Processing (NLP) that focuses on interpreting and extracting subjective information from textual data, typically categorizing it into positive, negative, or neutral sentiments. Over the past decade, sentiment analysis has gained significant attention due to its applications in various domains, including market research, social media monitoring, customer feedback management, and political analysis. This paper explores the methodologies used in sentiment analysis, from traditional machine learning approaches like Naive Bayes and Support Vector Machines to lexicon-based strategies and advanced deep learning models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers. It discusses the challenges of handling sarcasm, negation, and contextual nuances, as well as the diversity of languages in multilingual datasets. The paper highlights the impact of pre-trained language models like BERT and GPT, which have greatly improved sentiment classification accuracy, and examines practical applications in business and politics. Despite these advancements, there are ongoing challenges and research gaps, such as developing context-aware models and algorithms capable of understanding deeper semantic meanings and subtle emotional tones. This review provides a detailed overview of the current state of sentiment analysis, emphasizing both the progress made and the challenges that remain, and suggests potential future research directions in this rapidly evolving field..

**Keywords**: NLP, Machine learning, Customer feedback management

**Index**

|  |  |  |  |
| --- | --- | --- | --- |
| Index | | | Page no. |
| Chapter-1 | | |  |
|  | Introduction | |  |
|  |  |  |  |
| Chapter-2 | | |  |
|  | Literature Survey | |  |
|  | 2.1 | History |  |
|  | 2.1 | Review |  |
|  |  |  |  |
| Chapter-3 | | |  |
|  | Problem Statement | |  |
|  |  |  |  |
| Chapter-4 | | |  |
|  | Experimental Setup | |  |
|  | 4.1 | Hardware setup |  |
|  | 4.2 | Software Setup |  |
|  |  |  |  |
| Chapter-5 | | |  |
|  | Proposed system and Implementation | |  |
|  | 5.1 | Block Diagram of proposed system |  |
|  | 5.2 | Description of Block diagram |  |
|  | 5.3 | Implementation |  |
|  |  |  |  |
| Chapter-6 | | |  |
|  | Conclusion | |  |
|  |  |  |  |
| References | | |  |
|  |  |  |  |

# CHAPTER 1 INTRODUCTION

### INTRODUCTION

##### Sentiment analysis, or opinion mining, is a crucial aspect of Natural Language Processing (NLP) that aims to identify and interpret the emotional tone of text, categorizing it as positive, negative. This technique is widely utilized across various domains, including business, social media, healthcare, and politics, to provide valuable insights into public opinion and customer feedback. The process begins with text preprocessing, which involves cleaning and preparing the data through steps like tokenization, stop word removal, and stemming or lemmatization to make it suitable for analysis. Sentiment analysis can be approached using different methodologies: rule-based methods rely on predefined sentiment lexicons and pattern matching to identify sentiment based on specific words or syntactic structures, while machine learning techniques such as Naive Bayes, Support Vector Machines (SVM), and Logistic Regression use algorithms trained on labeled datasets to generalize from examples to unseen text. More recently, deep learning methods have transformed the field by employing advanced models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Transformers like BERT and GPT, which excel at capturing complex patterns and nuances in text data. Despite these advancements, challenges remain, including the detection of sarcasm and irony, understanding context-dependent meanings, handling negations, and analyzing multilingual data. Sarcasm and irony can obscure the true sentiment of a text, requiring models to grasp context and tone accurately. Contextual understanding is essential, as the meaning of words can change based on surrounding text, and handling negations is crucial for correct sentiment classification.

##### Moreover, the diversity of languages and cultural contexts adds complexity, necessitating models that can effectively analyze sentiment across multiple languages. Practical applications of sentiment analysis are vast, from monitoring customer feedback and tracking public opinion on social media to enhancing patient care and assessing political sentiments. Ongoing research aims to address these challenges by developing more sophisticated models that better understand language subtleties and multilingual data, ensuring sentiment analysis remains a powerful tool for extracting actionable insights from text data and supporting informed decision-making across various fields.

# CHAPTER 2 LITERATURE SURVEY

#### LITERATURE SURVEY

###### 2.1-HISTORY

Sentiment analysis has evolved significantly since its inception in the early 2000s, beginning with basic rule-based methods and machine learning techniques such as Naive Bayes and Support Vector Machines (SVM). These early approaches relied heavily on manually curated lexicons of sentiment-laden words and simple algorithms to classify text as positive, negative, or neutral. This initial work, led by pioneering researchers like Bo Pang and Lillian Lee, laid the groundwork for understanding how subjective information could be extracted from text. As the field developed, the rise of social media and the proliferation of user-generated content brought new challenges and opportunities. The massive volume of data generated on platforms like Twitter, Facebook, and review sites required sentiment analysis methods to adapt quickly. This period saw a shift towards more sophisticated models that could handle the intricacies of informal language, slang, emojis, and rapidly evolving internet vernacular. Researchers began developing more advanced techniques to detect specific emotions and opinion targets, moving beyond simple binary sentiment classification to a more nuanced understanding of text. The introduction of deep learning models, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), marked a turning point for sentiment analysis by allowing models to capture complex patterns and contextual information in ways that traditional methods could not. Transformers like BERT and GPT further revolutionized the field by enabling even deeper understanding and more accurate sentiment detection across a wide range of applications and contexts.

In recent years, sentiment analysis has continued to advance with the integration of pre-trained language models and the exploration of multimodal sentiment analysis, which combines text with images and videos for a more comprehensive understanding of sentiment. Researchers are also focusing on improving ethical considerations in sentiment analysis, ensuring fairness and transparency, especially in sensitive applications like hiring and law enforcement. Despite ongoing challenges, such as detecting sarcasm and processing multilingual data, sentiment analysis remains a powerful tool for extracting insights from text data, playing a critical role in areas like market research, social media monitoring, and customer feedback analysis.

**2.2 - LITERATURE REVIEW**

**"Sentiment Analysis of COVID-19 Related Tweets Using Transformer Models (IEEEAccess 2021) Jianqiang Huang, Xiaoyan Gu, Haiping Lu**

This study explores the use of transformer models, particularly BERT (Bidirectional Encoder Representations from Transformers), for sentiment analysis of tweets related to COVID-19. The authors utilize a large dataset of tweets to train and fine-tune the BERT model, achieving high accuracy in classifying sentiments as positive, negative, or neutral. This work demonstrates the effectiveness of transformers in handling large-scale, real-time data and their superior performance in understanding the nuances and context within social media text, especially during dynamic global events like the pandemic.

**"Multimodal Sentiment Analysis Using Text, Audio, and Visual Features" (IEEE Transactions on Affective Computing 2022) Carlos Busso, Mona Jalal, Murtaza Bulut**

Busso et al. present a comprehensive approach to sentiment analysis by integrating multiple modalities: text, audio, and visual data. The authors propose a deep learning framework that fuses these modalities to enhance sentiment classification accuracy. The study highlights the importance of considering different forms of data to capture a broader range of emotional expressions and sentiments. This multimodal approach is particularly useful in applications such as video sentiment analysis, where non-verbal cues play a critical role in conveying sentiment.

**"Zero-Shot Cross-Lingual Transfer for Sentiment Analysis" (ACL 2021) Yuan Zhang, Jieyu Zhang, Eduard Hovy**

This paper investigates the use of zero-shot cross-lingual transfer learning for sentiment analysis, aiming to reduce the dependency on large labeled datasets for every language. The authors employ a multilingual BERT model trained on multiple languages to perform sentiment analysis on languages that were not present in the training data. Their findings show that this method effectively transfers knowledge across languages, allowing for robust sentiment analysis in resource-poor languages without the need for extensive labeled data, thus advancing the field toward more inclusive and accessible NLP tools.

**"Explainable AI for Sentiment Analysis: A Survey" (IEEE Access 2022) Muhammad Atif Qureshi, Muhammad Usman Akram, Muhammad Atif Mughal**

Qureshi et al. provide a comprehensive survey on the application of explainable AI (XAI) in sentiment analysis, emphasizing the need for transparency and interpretability in models, particularly in sensitive areas like healthcare and finance. The paper reviews various techniques for explaining model decisions, such as feature importance scoring, saliency maps, and attention visualization, and discusses their effectiveness in sentiment analysis. This work highlights the growing demand for ethical AI practices and the importance of understanding how models derive their conclusions, ensuring fairness and accountability in sentiment analysis applications.

**"Aspect-Based Sentiment Analysis Using Graph Neural Networks" (EMNLP 2021) Ying Zhang, Jian Wu, Hua Xu**

This research focuses on aspect-based sentiment analysis (ABSA) using graph neural networks (GNNs), a relatively novel approach in the field. Zhang et al. propose a model that constructs a graph representing the relationships between different aspects of a product or service mentioned in text and uses GNNs to capture the dependencies and context around these aspects. This approach allows for more precise sentiment classification at the aspect level, significantly improving the model's ability to discern sentiment about specific attributes or features mentioned in reviews.

**"Improving Sentiment Analysis with Pre-trained Language Models: BERT and Beyond" (IEEE Intelligent Systems 2023) Rajesh Gupta, Sneha Sharma, Arjun Kumar**

Gupta et al. explore the advancements in sentiment analysis brought by pre-trained language models, such as BERT, RoBERTa, and ALBERT. The paper provides an empirical comparison of these models and their variations in sentiment analysis tasks across different datasets and domains. The study demonstrates that fine-tuning these models on domain-specific data can yield significant performance improvements, suggesting that future sentiment analysis research should focus on developing domain-adaptive models and leveraging transfer learning to enhance accuracy and generalization.

# CHAPTER 3

# Problem Statement

#### PROBLEM STATEMENT

Sentiment analysis, a subfield of Natural Language Processing (NLP), aims to automatically determine the sentiment or emotional tone expressed in text data, classifying it into categories such as positive, negative. Despite significant advancements in the field, several challenges continue to hinder the accuracy and reliability of sentiment analysis models. These challenges include the detection of sarcasm and irony, handling contextual nuances and negations, and managing the diversity and complexity of languages, especially in multilingual and code-mixed texts. Moreover, existing models often struggle to accurately capture sentiments in domain-specific contexts without extensive labeled datasets, limiting their applicability across different industries and languages. Additionally, there is a growing demand for explainable and transparent models that can provide insights into the decision-making processes of sentiment analysis algorithms, particularly in sensitive areas like healthcare, finance, and social media monitoring. Therefore, the problem at hand is to develop advanced sentiment analysis techniques that can address these challenges, improve model accuracy and robustness, and provide explainability, thereby enhancing their utility and applicability across diverse domains and languages.

# CHAPTER 4

# EXPERIMENTAL SETUP

**4. Experimental Setup**

**4.1 Hardware Setup :-**

* **CPU:** Multi-core processor (e.g., Intel i5/i7 or AMD Ryzen 5/7)
* **GPU:** Dedicated GPU (e.g., NVIDIA GTX 1660 or RTX 3060)
* **RAM:** Minimum 16 GB (32 GB recommended)
* **Storage:** SSD (512 GB or larger)
* **Network:** Stable internet connection
* **OS:** Windows

#### 4.2 Software Setup :-

#### Python: Version- 3.13

#### Jupyter Notebook : Package- notebook

#### NumPy: Package- numpy (latest: 1.25.0)

#### Package: pandas (latest: 2.0.3)

#### Matplotlib: Package- matplotlib (latest: 3.7.1)

#### Seaborn: Package- seaborn (latest: 0.12.2)

#### NLTK: Package- nltk (latest: 3.8.1)

#### Scikit-learn: Package- scikit-learn (latest: 1.3.0)

#### XGBoost: Package- xgboost (latest: 1.7.6)

#### WordCloud: Package- wordcloud (latest: 1.8.2)

#### Pickle: Built-in library (no package needed)

#### Regular Expressions (re): Built-in library (no package needed)

# CHAPTER 5

# Proposed System & Implementation

#### 5.Proposed system & Implementation

#### 5.1 Block diagram of proposed system:-

#### A diagram of a data analysis process Description automatically generated

#### 5.2 Description of block diagram :-

#### Load Data and Clean: Load raw text data (reviews), remove noise (stop words, special characters),

#### and handle missing values.

#### Feature Engineering: Create additional features to improve the model, like word frequency

#### or sentiment-specific indicators.

#### Exploratory Data Analysis (EDA): Analyze the data visually (e.g., word clouds, sentiment distribution) to understand key patterns and correlations

#### Text Preprocessing: Prepare text by converting it into tokens, lowercasing, removing punctuation,

#### and applying techniques like stemming/lemmatization.

#### Feature Extraction: Convert text into numerical format using methods like TF-IDF, bag-of-words,

#### or word embeddings (e.g., Word2Vec).

#### Split and Scale Data: Split data into training and testing sets. Scale numeric features if necessary

#### for uniformity.

#### Model Building (XGBoost): Train an XGBoost model, which uses gradient boosting for efficient

#### and accurate sentiment classification.

#### Model Evaluation: Measure performance with accuracy, precision, recall, and confusion matrix

#### to evaluate predictions (e.g., positive/negative).

#### Results and Insights: Analyze model outputs, understand key features driving sentiment, and

#### visualize results for actionable insights.

#### 5.3 Implementation

#### 5.4 Advantages/ Application/ result table

|  |  |
| --- | --- |
| **Category** | Details |
| Advantages | - Real-time analysis: Helps in understanding customer sentiment in real-time. - Scalability: Can analyze large volumes of text automatically. - Actionable insights: Helps in deriving insights that can improve customer experience and business strategies. - Efficiency: Automates the process of reading and analyzing opinions, saving time and resources. - Customization: Models can be fine-tuned for specific languages, industries, or contexts. |
| Applications | - Customer feedback analysis: Used by companies to analyze product reviews, support tickets, or social media interactions to improve services. - Brand monitoring: Identifies positive or negative mentions about a brand on social platforms. - Market research: Helps companies understand public sentiment toward a product, service, or trend. - Financial market prediction: Used in stock price predictions by analyzing news sentiment. - Political analysis: Identifies voter sentiment towards political parties, candidates, or issues. |
| Results | - Improved decision-making: Companies can make data-driven decisions based on customer sentiments. - Enhanced customer satisfaction: Faster resolution of issues by identifying negative sentiment early. - Increased profitability: By analyzing feedback and improving products based on sentiment analysis. - Sentiment classification accuracy: Models (like XGBoost, SVM) typically achieve over 85% accuracy with properly preprocessed data. - Actionable business insights: Helps businesses react to trends and sentiments swiftly. |

# CHAPTER 6

# Conclusion

#### 6. Conclusion

#### In this project, sentiment analysis using natural language processing (NLP) was successfully implemented, demonstrating the power of machine learning algorithms, such as XGBoost, to classify and interpret human emotions from text data. The project showcased the importance of data cleaning, feature engineering, and text preprocessing techniques in building accurate models that can extract meaningful insights from vast amounts of unstructured data. The results highlighted the potential for sentiment analysis to assist businesses in making informed decisions by analyzing customer feedback, monitoring brand sentiment, and predicting market trends.

#### Overall, sentiment analysis using NLP provides an efficient and scalable approach to understanding public opinion, enabling companies to improve products and services and react proactively to changes in customer sentiment.

#### 7. Future Scope

#### Future work can focus on improving the accuracy and robustness of sentiment analysis models by:

#### Exploring advanced deep learning techniques like transformers (BERT, GPT) for better context understanding and handling sarcasm or complex emotions.

#### Multilingual sentiment analysis, allowing models to process feedback in various languages.

#### Aspect-based sentiment analysis, which can break down sentiment at a more granular level (e.g., sentiment toward specific product features).

#### Real-time sentiment monitoring using streaming data from social media platforms or customer service channels.

#### Hybrid models that combine sentiment analysis with other NLP tasks like topic modeling or intent detection for more comprehensive insights.