***Mini Project***

**Sentiment Analysis**

***Submitted to***

**Rashtrasant Tukdoji Maharaj Nagpur University**

****

***Submitted by***

**1) Anagha Borkar**

**2) Anjali Chaudhari**

**3) Bhumika Bagade**

**4) Divya Ikhar**

**5) Ekta Chopde**

**Under the guidance of**

**Guide**

**Dr. G. M Asutkar**

**Prof. Kaveri Deosarkar**

**DEPARTMENT OF ARTIFICIAL INTELLIGENCE & DATA SCIENCE**

**PRIYADARSHINI COLLEGE OF ENGINEERING, NAGPUR**

**SESSION 2023-2024**

(I)

**CERTIFICATE**

This is to certify that Anagha Borkar,Anjali Chaudhari, Bhumika Bagade, Divya Ikhar, Ekta Chopde have successfully completed the mini project entitled “Sentiment Analysis” as a part of B.Tech 6th Semester for the Academic Session 2022-2023 in the Department of Artificial Intelligence & Data Science, Priyadarshini College of Engineering, Nagpur.

The mini project work has been completed under our guidance and supervision, and to the best of our knowledge, this mini project report is the original work of the students. No part of this mini project report has been submitted for any other degree or diploma.

We hereby declare that the mini project work has been carried out by the students under our guidance and supervision.

Name and Signature of the Guide :

Dr. G. M Asutkar

Prof. Kaveri Deosarkar HoD

Department of Artificial

Intelligence & Data Science

Date: 05-APRIL-2024

Place: Department of Artificial Intelligence & Data Science, Priyadarshini College of Engineering, Nagpur.

(II)

**ACKNOWLEDGEMENT**

I would like to express my sincere gratitude to Dr. G. M Asutkar for providing me with the opportunity to carry out this mini project and for their valuable guidance and support throughout the mini project. Their insights and feedback were instrumental in shaping the direction and scope of the mini project.

I would also like to extend my thanks to Prof. Kaveri Deosarkar for their support and assistance in completing the project.

I am grateful to the Head of the department Dr. G. M. Asutkar and faculty members of the Department of Artificial Intelligence & Data Science for their valuable inputs and suggestions that helped me in completing the project.

I would like to express my thanks to the Principal Dr. S. A. Dhale & Vice Principal Dr. G. M. Asutkar of Priyadarshini College of Engineering, Nagpur for providing me with the necessary resources and facilities to carry out the mini project.

Finally, I would like to thank my family and friends for their constant encouragement and support throughout the project.

[1]Anagha Borkar

[2] Anjali Chaudhari

[3] Bhumika Bagade

[4] Divya Ikhar

[5] Ekta Chopde

DATE:-05-APRIL-2024

(III)

**ABBREVIATIONS**

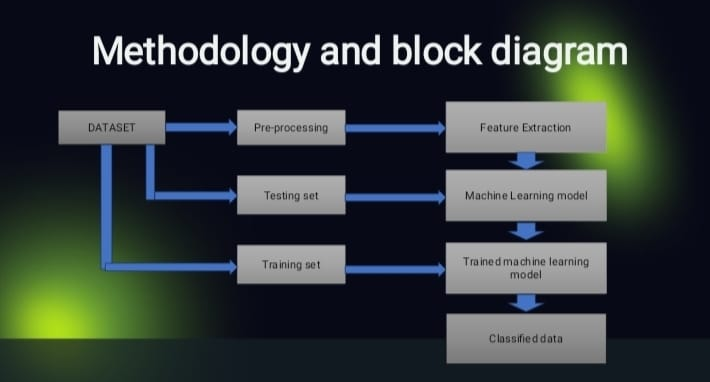
|  |  |  |
| --- | --- | --- |
| **Abbreviation** | **Meaning** | **Page No.** |
| pd | pandas | 1 |
| np | numpy | 1 |
| sns | seaborn | 1 |
| plt | Procedural linkage table | 4 |
| ML | Machine learning | 6 |
| SVM | Support vector machine | 6 |

(IV)

**INDEX**

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Name of Chapter/ Topic** | **Page No.** |
| **01** | **Introduction** | **1** |
| **02** | **Problem Statement** | **3** |
| **03** | **Literature Survey** | **5** |
| **04** | **Methodology** | **7** |
| **05** | **Result and Discussion** | **9** |
| **06** | **Conclusion** | **11** |
| **07** | **Future Scope** | **12** |
| **08** | **References** | **14** |

**INTRODUCTION**

****

***Fig. 1. Block Diagram of Sentiment Analysis***

Here’s a breakdown of the methodology:

* **Dataset:** This is the block where the raw data is stored.
* **Pre-processing:** This block is where the data is prepared for use in the machine learning model. This may involve cleaning the data, formatting the data, and handling missing values.
* **Feature extraction:** In this block, features are extracted from the data. Features are characteristics of the data that will be used by the machine learning model to make predictions.
* **Splitting the data into a training set and a testing set:** The data is split into two sets: a training set and a testing set. The training set is used to train the machine learning model, and the testing set is used to evaluate the performance of the model.
* **Machine learning model:** This block represents the machine learning model that is used to make predictions. The model is trained on the training set, and then it is used to make predictions on the testing set.
* **Training the machine learning model:** This block refers to the process of training the machine learning model on the training set. During training, the model learns the patterns in the data.
* **Classified data:** This block represents the output of the machine learning model. The model will take the data from the testing set and classify it according to the patterns it learned during training.

Block diagrams are a simplified way to represent complex systems. They are often used to communicate the overall flow of a process or system without getting bogged down in the details.

**CHAPTER - II**

**PROBLEM STATEMENT**

* **Objective**: Develop a sentiment analysis model to classify text data into positive, negative, or neutral sentiments.
* **Scope**: The project will focus on analyzing sentiments from user-generated text data such as product reviews, social media posts, or customer feedback.
* **Dataset**: Acquire a sizable dataset containing text samples labeled with corresponding sentiment categories (positive, negative, neutral).
* **Preprocessing**:
  + Clean the text data by removing special characters, punctuation, and irrelevant information like URLs or usernames.
  + Tokenize the text into words or subword units.
  + Handle issues like misspellings, contractions, and abbreviations.
  + Perform normalization techniques such as lowercasing or lemmatization.
* **Feature Extraction**:
  + Convert text data into numerical vectors suitable for machine learning algorithms.
  + Explore techniques like Bag-of-Words, TF-IDF, or word embeddings (e.g., Word2Vec, GloVe) to represent textual features.
* **Model Selection**:
  + Experiment with various machine learning models such as Naive Bayes, Support Vector Machines (SVM), Logistic Regression, or deep learning architectures like Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), or Transformer-based models (e.g., BERT, GPT).
  + Evaluate each model's performance using appropriate metrics like accuracy, precision, recall, and F1-score.
* **Hyperparameter Tuning**:
  + Fine-tune the selected model's hyperparameters using techniques like grid search, random search, or Bayesian optimization to improve performance.
* **Training and Validation**:
  + Split the dataset into training, validation, and test sets to train and validate the model's performance.
  + Employ techniques like cross-validation to ensure robustness and generalizability of the model.
* **Evaluation**:
  + Evaluate the model's performance on the test set using appropriate evaluation metrics.
  + Analyze the confusion matrix to understand the model's behavior concerning different sentiment classes.
* **Deployment**:
  + Once satisfied with the model's performance, deploy it in a real-world application environment.
  + Implement monitoring mechanisms to track the model's performance over time and retrain it if necessary.
* **Continuous Improvement**:
  + Gather feedback from users and monitor the model's performance in the production environment.
  + Periodically retrain the model with new data to adapt to evolving language patterns and sentiment trends.
* **Documentation and Reporting**:
  + Document the entire process including data collection, preprocessing steps, model architecture, training methodology, and evaluation results.

**CHAPTER-III**

**LITERATURE SURVEY**

- **Research Papers**:- Review academic papers from various journals and conferences specializing in sentiment analysis.

- Look for papers discussing different approaches such as lexicon-based methods, machine learning algorithms, deep learning models, and hybrid approaches.

- Analyze papers that delve into evaluation metrics used to measure the performance of sentiment analysis models.

- Identify studies focusing on specific domains like social media, product reviews, or movie reviews to understand domain-specific challenges and techniques.

- **Survey Papers**:- Explore survey papers that provide comprehensive summaries of sentiment analysis techniques, trends, and challenges.

- Look for papers that cover recent advancements in sentiment analysis methodologies and their applications across different domains.

- Analyze survey papers to identify gaps in existing research and potential directions for future exploration.

- Pay attention to surveys that discuss the limitations of current sentiment analysis techniques and propose novel solutions or improvements.

- **Datasets**:- Investigate commonly used datasets for sentiment analysis tasks, such as the IMDb movie review dataset, the Yelp review dataset, or the Twitter sentiment dataset.

- Understand the characteristics of each dataset, including the size, diversity of samples, and annotation quality.

- Consider datasets annotated with fine-grained sentiment labels (e.g., positive, negative, neutral) for more nuanced sentiment analysis tasks.

- Explore domain-specific datasets tailored to specific applications like e-commerce, healthcare, or finance to address domain-specific challenges.

- **Methodologies**:- Analyze different methodologies and techniques employed in sentiment analysis, including rule-based methods, machine learning algorithms (e.g., Naive Bayes, Support Vector Machines), deep learning models (e.g., Recurrent Neural Networks, Transformers), and hybrid approaches.

- Explore the strengths and weaknesses of each methodology in terms of accuracy, scalability, interpretability, and computational efficiency.

- Consider the suitability of each methodology for specific tasks, such as sentiment polarity classification, aspect-based sentiment analysis, or sentiment intensity prediction.

- Investigate recent advancements in sentiment analysis techniques, such as transfer learning, self-supervised learning, or multimodal sentiment analysis.

- **Evaluation Metrics**:- Examine various evaluation metrics used to assess the performance of sentiment analysis models, including accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC).

- Understand the characteristics of each evaluation metric and their relevance to specific sentiment analysis tasks.

- Consider additional metrics like mean squared error (MSE) for regression-based sentiment analysis tasks or Cohen's Kappa coefficient for inter-annotator agreement analysis.

- Explore techniques for evaluating sentiment analysis models in domain adaptation or cross-domain sentiment analysis scenarios.

**CHAPTER - IV**

**METHODOLOGY**

- **Define Objectives**:- Clearly defining objectives involves outlining the purpose of the sentiment analysis project. For instance, it could be to gauge public opinion about a particular product, service, or event, or to understand sentiment trends over time. Specific questions could include:

- What is the overall sentiment towards our brand/product?

- What are the main topics driving positive/negative sentiment?

- How does sentiment vary across different demographics or geographic regions?

- **Data Collection**:- Gathering data involves identifying and collecting sources that contain relevant text data. This could include:

- Social media platforms like Twitter, Facebook, or Reddit.

- E-commerce websites for product reviews.

- Customer feedback forms or surveys.

- News articles or blog posts related to the topic of interest.

- It's essential to ensure that the data collected is representative and covers the range of sentiments you want to analyze.

- **Preprocessing**: - Cleaning and preprocessing the data involves preparing it for analysis by removing noise and irrelevant information. This typically includes:

- Removing special characters, punctuation, and HTML tags.

- Handling missing values through imputation or removal.

- Tokenization: splitting text into individual words or tokens.

- Removing stopwords: common words like "and," "the," etc., that don't contribute much to sentiment analysis.

- Stemming or lemmatization: reducing words to their base or root form to normalize the data.

- **Labeling (Optional):**- Labeling the data is necessary for supervised learning approaches, where the model needs to be trained on labeled examples. This involves manually assigning sentiment labels (positive, negative, neutral) to a portion of the text data. Labeling can be done through manual annotation or crowdsourcing platforms.

- **Feature Engineering:**- Feature engineering is the process of transforming raw text data into numerical features that can be used as input to machine learning algorithms. Some common techniques include:

- Word frequency: counting the occurrence of each word in the text.

- N-grams: sequences of adjacent words of length n.

- Part-of-speech (POS) tags: categorizing words based on their grammatical roles.

- Sentiment lexicons: lists of words annotated with sentiment scores.

- Word embeddings: dense vector representations of words learned from large text corpora using techniques like Word2Vec or GloVe.

- These features help capture the semantic and syntactic information present in the text, enabling the model to learn patterns related to sentiment.

**CHAPTER - V**

**RESULTS AND DISCUSSION**

- **Result Interpretation:**

- Determine the overall sentiment expressed in the text data: positive, negative, or neutral.

- **Discussion Implications:**

- Relate sentiment results to the objectives of the analysis.

- Discuss patterns or trends observed in the sentiment analysis.

- Address any limitations encountered during the analysis.

- Suggest potential areas for further investigation or improvement.

- **Comparison and Context:**

- Compare sentiment analysis results with other analyses or previous studies if available.

- Provide context for the findings within the subject matter or application domain.

- **Broader Implications**:

- Consider the broader implications of the sentiment analysis results.

- Discuss how the findings can enhance customer engagement and inform targeted marketing strategies.

**Elaboration:**

**1.** **Result Interpretation:**

- The sentiment analysis results provide insights into the prevailing sentiment within the analyzed text data. By identifying whether the sentiment is positive, negative, or neutral, businesses can gauge the overall perception or attitude towards their products, services, or brand.

**2. Discussion Implications:**

- Relating the sentiment results to the objectives of the analysis helps in understanding how well the sentiment aligns with the desired goals. Discussing patterns or trends observed in sentiment can uncover valuable insights into customer preferences, pain points, or satisfaction levels. Addressing limitations encountered during the analysis ensures transparency and helps in accurately interpreting the findings. Suggesting potential areas for further investigation or improvement guides future research or strategy development.

**3. Comparison and Context:**

- Comparing sentiment analysis results with other analyses or previous studies provides a broader perspective and validates the findings. Providing context for the sentiment findings within the subject matter or application domain helps stakeholders understand the relevance and significance of the results in real-world scenarios.

**4. Broader Implications:**

- Considering the broader implications of the sentiment analysis results involves exploring how the insights gained can drive enhanced customer engagement and inform targeted marketing strategies. By understanding customer sentiment, businesses can tailor their marketing efforts, messaging, and product offerings to better resonate with their target audience, ultimately improving customer satisfaction and loyalty.

**CHAPTER - VI**

**CONCLUSION**

- **Big Data Opportunities for Sentiment Analysis**:- Big data and related technologies provide vast amounts of data for sentiment analysis.

- Enables analysis of large-scale datasets from various sources like social media, reviews, and surveys.

- Offers potential for more comprehensive insights into public opinion

and sentiment trends.

**- Challenges with Big Data Quality:**- Big data quality issues such as noise, inconsistencies, and inaccuracies.

- Poor data quality can significantly impact the accuracy and reliability of sentiment analysis results.

- Challenges in ensuring data cleanliness, consistency, and relevance.

**- Gap in Current Sentiment Analysis Approaches:**- State-of-the-art sentiment analysis approaches often overlook the issue of big data quality.

- Existing methods may focus more on algorithmic advancements rather than data quality considerations.

- Lack of integration of data quality assurance mechanisms into sentiment analysis frameworks**.**

**- Importance of Addressing Big Data Quality:**- Poor data quality can lead to misleading or erroneous sentiment analysis outcomes.

- Neglecting data quality can undermine the credibility and usefulness of sentiment analysis results.

- Ensuring data quality is crucial for deriving actionable insights and making informed decisions.

**CHAPTER - VII**

**FUTURE SCOPE**

1. What is the Relevance of your Project?

The relevance of sentiment analysis projects lies in their ability to extract insights from vast amounts of textual data, helping businesses and organizations understand public opinion, customer feedback

1. What is the Novelty in your project?

Data Source, Model Architecture, Feature Engineering, Multimodal Sentiment Analysis

1. What is the proposed Major Project Title?

Enhancing Sentiment Analysis Accuracy and Efficiency through Deep Learning and Multimodal Fusion Techniques

1. What is the Problem statement?

In a sentiment analysis project, the problem statement typically outlines the goal of the project, which is to analyze and understand the sentiment expressed in text data.

1. Is it Industry Relevant?

Yes, sentiment analysis is highly relevant in various industries including marketing, customer

service, finance, social media monitoring, and market research.

6. Can it be Sponsored?

Yes, sentiment analysis projects can be sponsored by companies, research institutions, or

government agencies

1. Is there any Relevance with India AI web. Yes/No?

yes

If Yes,

1. What is the topic?

Sentiment analysis

1. What are the Requirements?

I. Software: machine learning, logistic regression, NLP, confusion matrix

1. What is the timeline for completion?

15 day

**CHAPTER - VIII**

**REFERENCES**

[1] Ancy Grana, Sentiment Analysis of text using machine learning models

, May 2022

[2] Anjali Gupta, Sentiment Analysis using machine learning , September 2021

[3] Sean Humpherys, Using machine learning sentiment analysis to evaluate learning impact*,*

February 2022

[4] Dipak Kawade, *Sentiment analysis :* Machine learning approach*, July 2019*