|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **PROJECT OVERVIEW STATEMENT** | **Project Name:**  Sentiment Analysis of Tweets Using Machine Learning | | **Student Name:**  Yeshwanth Gujjalapati | |
| **Problem/Opportunity:** | | | | |
| People share their opinions on social media platforms like Twitter. Understanding whether tweets are positive, or negative is useful for businesses, politics, and customer service. Manually reading millions of tweets is impossible, so machine learning can help classify tweets automatically. This project aims to build a system that can automatically detect if a tweet is positive or negative. | | | | |
| **Goal:** | | | | |
| The goal of this project is to build a sentiment analysis system that can classify tweets as positive or negative using machine learning. The system should achieve at least 70% accuracy within 10 weeks.   * **Specific**: The project will develop a machine-learning pipeline to classify tweets. * **Measurable**: The system should achieve at least 70% accuracy on the test dataset. * **Assignable**: One student (you) will handle data preparation, model training, and testing. * **Realistic**: The project will use standard Python libraries and a publicly available dataset. * **Time-Related**: The project will be completed in 10 weeks, with model testing and evaluation in the last week. | | | | |
| **Objectives:** | | | | |
| 1. **Prepare and Clean the Twitter Dataset**    1. Outcome: Have a dataset free from unnecessary elements like URLs, mentions, and special characters.    2. Time Frame: Completed by the end of Week 3.    3. Measure: The cleaned dataset should be ready for training the models.    4. Action: Use Python scripts to remove noise and preprocess text. 2. **Train Machine Learning Models**    1. Outcome: Set up and train at least two machine learning models (Logistic Regression, Random Forest).    2. Time Frame: Completed by the end of Week 6.    3. Measure: Models should be running without errors, and initial results should be available.    4. Action: Use scikit-learn or similar libraries for training. 3. **Test and Evaluate Model Performance**    1. Outcome: Analyze performance using accuracy, precision, recall, F1-score, and confusion matrix.    2. Time Frame: Completed by the end of Week 8.    3. Measure: Compare performance with the 70% accuracy goal.    4. Action: Tune model parameters and compare results. 4. **Complete Report and Presentation**    1. Outcome: Submit a full report and demonstrate the system.    2. Time Frame: Completed by the end of Week 10.    3. Measure: Submit a detailed report and show how the system works.    4. Action: Summarize results, create visualizations, and present findings. | | | | |
| **Success Criteria:** | | | | |
| * The model should achieve at least 70% accuracy. * All project steps (data preparation, model training, testing, and report) should be finished on time. * The system should correctly classify tweets into positive or negative categories. * The final report and demonstration should clearly explain the system and its results. | | | | |
| **Assumptions, Risks, Obstacles:** | | | | |
| * Dataset Availability: The dataset should be reliable and accessible. * Computational Power: Running models on a large dataset may be slow on limited hardware. * Time Constraints: Any data cleaning or training delays could affect deadlines. * Model Performance: The model may not work well on new, unseen tweets without diverse training data. * Resource Availability: The project needs a good computer or cloud environment for processing large data. | | | | |
| **Prepared By** | **Date** | **Approved By** | | **Date** |
| Yeshwanth Gujjalapati | 13th Feb, 2025 |  | |  |