|  |  |
| --- | --- |
| Internship Project Title | TCS iON RIO 125 - Automate sentiment analysis of textual comments and feedback |
| Project Title | Twitter Sentiment Analysis |
| Name of the Company | TCS iON |
| Name of the Industry Mentor | Debashis Roy |
| Name of the Institute | Vishwakarma University, Pune |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Start Date | End Date | Total Effort (hrs.) | Project Environment | Tools used | |
| 10/12/2023 | 08/01/2024 | 126 | Python | | Jupiter Notebook, Excel |
| Project Synopsis:  The project aims to perform real-time sentiment analysis on Twitter data, utilizing advanced natural language processing techniques. Through the analysis of tweets, the system classifies sentiments as positive, negative, or neutral, offering valuable insights into the collective emotional tone of the Twitter community.  Objective:  The main objective of the project is to develop a real-time sentiment analysis system for Twitter data, employing advanced natural language processing and machine learning algorithms. The system aims to classify tweets into positive, negative, or neutral sentiments, providing valuable insights into the evolving trends of public opinion on the platform. Through a user-friendly interface and robust handling of noisy data, the project seeks to offer stakeholders a scalable and efficient tool for making informed decisions based on real-time sentiment analysis.  Dataset Description:  The dataset provided is the **Sentiment140 Dataset**which consists of **1,600,000 tweets** that have been extracted using the Twitter API. The various columns present in this Twitter data are:   * **target:**the polarity of the tweet (positive or negative) * **ids:**Unique id of the tweet * **date:**the date of the tweet * **flag:**It refers to the query. If no such query exists, then it is NO QUERY. * **user:** It refers to the name of the user that tweeted * **text:** It refers to the text of the tweet   Methodology:   1. Data Exploration: Conduct exploratory data analysis (EDA) to understand the distribution of sentiments, text lengths, and potential trends over time. Visualize key insights using charts and graphs. 2. Preprocessing: Clean and preprocess the textual data by removing stopwords, handling punctuation, and applying techniques like tokenization and lemmatization. 3. Sentiment Classification: Develop a sentiment analysis model using natural language processing (NLP) techniques. Train the model to classify individual reviews into positive, negative, or neutral sentiments. 4. Overall Sentiment Prediction: Extend the model to predict the overall sentiment of products by considering the aggregated sentiments within their reviews. Address challenges associated with contradictory statements. 5. Model Evaluation: Assess the accuracy and performance of the sentiment analysis model using appropriate metrics. Tackle issues such as imbalanced datasets to ensure robust generalization. 6. Business Insights: Translate sentiment analysis results into actionable insights for businesses. Provide strategic recommendations based on identified strengths, weaknesses, and areas for improvement.   Expected Outcomes:  The anticipated outcomes of this project encompass several key aspects. First and foremost, the system will be geared towards classifying tweets in real-time, assigning them into positive, negative, or neutral sentiment categories. This dynamic classification process will occur as new tweets are continuously fetched from Twitter. Furthermore, the project will deliver visual representations, such as charts or graphs, showcasing the changing trends in sentiments on Twitter. Users can glean valuable insights into the prevailing emotions within the Twitter community through these visualizations. Additionally, the system aims to provide accessible and user-friendly interfaces, making it easier for stakeholders to interact with and interpret sentiment analysis results. Robustness is a priority, ensuring the system can adeptly handle noisy data, including misspellings or slang often present in tweets. Overall, the expected outcome is a reliable and scalable system that not only informs stakeholders about current sentiments on Twitter but also offers potential pathways for future enhancements, such as incorporating sentiment intensity analysis or adapting to evolving language trends. | | | | | |
| Solution Approach:   * Outline the step-by-step approach you followed in your project, starting from data loading and preprocessing to feature engineering, model fitting, and evaluation. * Emphasize the use of exploratory data analysis (EDA) and various visualizations for a comprehensive understanding of the dataset. * Specify the text processing techniques applied, such as cleaning, stopwords removal, stemming, and lemmatization. * Detail the feature engineering steps, including the creation of TF-IDF matrices and data splitting. * Highlight the oversampling technique employed for addressing data imbalances. * Clearly articulate the choice of classifiers (Naive Bayes and XGBoost) and the evaluation metrics used (F1 score). | | | | | |
| Assumptions:  In the pursuit of developing a real-time sentiment analysis system for Twitter data, certain assumptions are made to guide the project's scope and implementation. Firstly, it is assumed that the Twitter API provides consistent and reliable access to real-time data, enabling the continuous collection of tweets for analysis. Additionally, the project operates under the assumption that the textual content of tweets alone is sufficient for accurate sentiment analysis, without incorporating multimedia elements such as images or videos. Furthermore, the effectiveness of the sentiment analysis model relies on the assumption that the labeled training data used for machine learning algorithms adequately represents the diverse range of sentiments expressed on Twitter. These assumptions serve as foundational considerations for the project, shaping the approach and providing a framework for development. Regular evaluation and validation against these assumptions will be integral to ensuring the robustness and reliability of the sentiment analysis system throughout its implementation.Top of Form | | | | | |
| Project Diagrams:  class imbalance twitter sentiment analysis  target class  wordcloud twitter sentiment analysis    wordcloud  ***Model-1:***  model 1 evaluation twitter sentiment analysis  ***ROC-AUC Curve for model-1 :***  roc curve | | | | | |
| ***Model-2:***  evaluation model 2 twitter sentiment analysis  ***ROC-AUC Curve for model-2 :***  roc  ***Model-3:***  model 3  ***ROC-AUC Curve for model-3 :***  roc 3 twitter sentiment analysis  Conclusion:  Upon evaluating all the models, we can conclude the following details i.e.  **Accuracy:** As far as the accuracy of the model is concerned, Logistic Regression performs better than SVM, which in turn performs better than Bernoulli Naive Bayes.  **F1-score:** The F1 Scores for class 0 and class 1 are : (a) For class 0: Bernoulli Naive Bayes(accuracy = 0.90) < SVM (accuracy =0.91) < Logistic Regression (accuracy = 0.92) (b) For class 1: Bernoulli Naive Bayes (accuracy = 0.66) < SVM (accuracy = 0.68) < Logistic Regression (accuracy = 0.69)  **AUC Score:** All three models have the same ROC-AUC score.  We, therefore, conclude that the Logistic Regression is the best model for the above-given dataset. | | | | | |
| Algorithms:  For the sentiment analysis project focused on real-time Twitter data, a judicious combination of the Bernoulli Naive Bayes classifier and Logistic Regression is proposed. The Bernoulli Naive Bayes classifier will be instrumental in capturing the presence or absence of key words in tweets, making it apt for binary sentiment classification prevalent in social media contexts. This algorithm excels in simplicity and computational efficiency, crucial for real-time analysis on a platform with dynamic content. In tandem, Logistic Regression will be employed to estimate the probability of tweets belonging to specific sentiment classes. Known for its interpretability, Logistic Regression provides insights into the influence of individual features on sentiment, offering a nuanced understanding of sentiment dynamics. This amalgamation seeks to strike a balance between effectiveness and interpretability, addressing the specific challenges presented by sentiment analysis on Twitter data. The combined approach will undergo iterative refinement and validation against real-world data to ensure optimal performance in capturing the evolving sentiments expressed on the platform.  Top of Form | | | | | |
| Outcome:  The goal of our project is to create a smart tool for understanding how people feel on Twitter in real-time. We want this tool to be really good at telling if tweets are positive, negative, or just neutral. Imagine it like a friendly assistant that can quickly show you how the mood on Twitter is changing. We're also making sure it's easy for everyone to use and that it works well even when people use funny or misspelled words. Our aim is to continuously make it better based on feedback and to help businesses and researchers understand people's feelings on Twitter more easily. The application of the Naive Bayes model and XGBoost for sentiment classification demonstrated the project's success in predicting sentiments accurately. The consideration of the F1 score as an evaluation metric ensured a balanced assessment of the models' performance. | | | | | |
| Exceptions considered:  In developing the sentiment analysis system for real-time Twitter data, several exceptions have been carefully considered to ensure the robustness and accuracy of the results. One critical consideration involves the handling of noisy data, encompassing issues such as misspellings, slang, and ambiguous language commonly found in tweets. To address this, the system incorporates robust preprocessing techniques to clean and normalize the text, minimizing the impact of irregularities on sentiment analysis. Additionally, the project acknowledges the potential presence of biases in the training data and aims to account for these biases to avoid skewed sentiment predictions. The system is designed to be resilient in the face of ambiguous or sarcastic language, acknowledging that these linguistic nuances can significantly influence sentiment interpretation. By taking these exceptions into account, the project aims to enhance the system's adaptability to the diverse and dynamic language patterns inherent in Twitter data, ultimately improving the accuracy and reliability of sentiment analysis results. | | | | | |
| Enhancement Scope:  We have exciting plans to make our sentiment analysis tool even better! First, we want to understand how strongly people feel, not just if they feel positive or negative. We're also working on making sure our tool stays up-to-date with the latest words and phrases people use on Twitter. If you're interested in where tweets are coming from, we're exploring ways to show how feelings vary in different places. We're thinking about including pictures and videos in our analysis too, so we get the full picture of what people are saying. Your feedback is important too, so we're working on a way for you to tell us how accurate our tool is. And there's more – we're looking into showing not just overall feelings but also feelings about specific topics or subjects. It's like giving our tool superpowers to understand the ever-changing language of Twitter and make it even more helpful for all! Top of Form | | | | | |
| Link to Code and executable file:  <https://github.com/BandalChintamani99/Sentiment-analysis-on-Twitter-Data/commit/66c3d71f2d51dc4152bb122080a96894dcf88043?diff=unified&w=0> | | | | | |