**EVALUATION**

This chapter focuses on evaluating the Tweets2GPT project, which is designed to improve Thameslink train maintenance through Natural Language Processing (NLP) Analysis. The chapter provides an evaluation of the methods used to evaluate the project's effectiveness.

***Purpose of Evaluation***

Evaluating the Tweets2GPT project is important to understand its impact on improving Thameslink train maintenance through NLP Analysis. During the evaluation phase, we aim to identify the strengths, weaknesses, and areas of improvement of the project.

***The Importance of Evaluation***

Evaluation acts as a compass and draws a road map for The Fine Tuners team to identify and eliminate shortcomings. By both filling gaps and ensuring sustainability, evaluation helps the project achieve its goals by shaping next steps.

***Evaluation Process***

A step-by-step approach was followed during the evaluation phase. The Evaluation phase started by examining Irony, and then moved on to Sentiment Analysis and LLM Analysis. Finally, the evaluation was concluded by controlling how the project affected the company's Key Performance Indicators (KPIs). This sequence helped make the impact of the project manageable.

***The Process of Modelling and Evaluation***

The diagram below shows how both the model and the evaluation process work collaboratively. The visual presentation highlights key steps in the methodology and outlines the step-by-step workflow. This visual gives an idea of the functioning of the model and subsequent evaluation procedures.

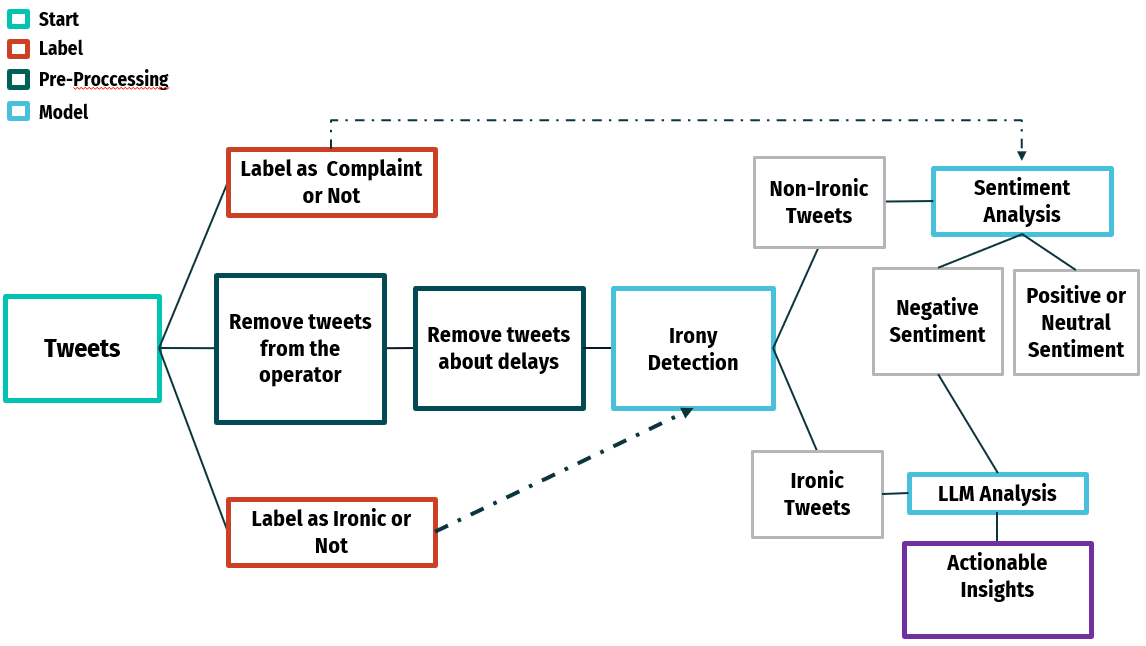


Illustration 1- Process of Modelling and Evaluation

The evaluation started with two separate surveys dedicated to the analysis of Irony and Sentiment. All team members responded to surveys categorized as Ironic or Non-Ironic and Complaint or Non-Complaint. The answered questions consisted of 100 randomly selected tweets. The results of these surveys were set aside for further evaluation steps.

The process was simplified by initially filtering tweets to select only those relevant to the maintenance team. The system eliminated tweets from Thameslink and those including delays, focusing instead on tweets that pointed out specific issues. Following this, irony detection was applied to these tweets. The model then identified ironic tweets and directed them to LLM Analysis. The non-ironic tweets underwent Sentiment Analysis; those classified as negative were also sent to LLM Analysis. Positive and neutral tweets, deemed irrelevant for the maintenance team, were disregarded. Tweets that exhibited both irony and negative sentiments were particularly valuable during the LLM analysis, providing significant insights for the maintenance team's consideration.

For the Irony and Sentiment Evaluation, a similar filtering process was applied to the 100 random tweets as done in the model. The results were then compared to evaluate the accuracy of the model’s classifications against the team's assessments from the surveys. In the LLM Evaluation, the focus was on assessing the consistency of the LLM Analysis by comparing its outputs for identical inputs. A more detailed explanation of these processes will be presented in the following sections of the report.

**Evaluation of Irony**

***Defining Irony***

Irony, as per the Oxford Dictionary, is the use of words that convey the opposite of their literal meaning, often in a humorous or sarcastic manner.

***Significance of Identifying Irony***

Understanding and identifying ironic tweets is crucial for:

* **Accurate Sentiment Analysis:** It is crucial for understanding the true sentiment in passenger messages, avoiding oversight of complaints or problems.
* **Preventing Misinterpretation:** It prevents misinterpretation during the analysis of passenger opinions, contributing to a clearer understanding and more effective responses.
* **Enabling Effective Responses:** It enables a clear understanding for effective responses, ensuring that actions are based on an accurate interpretation of the underlying sentiment.

***Importance of Irony Detection***

The utilization of "irony detection" in the process is crucial to mitigate misleading analysis. The primary purpose is to mark ironic tweets and alert the LLM Analysis at the end of the pipeline. This precaution ensures that ironic tweets do not lead to misleading conclusions during the analysis.

***Evaluation of Irony Detection***

To assess how well the "irony detection" works, a two-step method is used. Firstly, the outcomes of the model are compared with insights from a survey where all group members give their opinions, bringing different viewpoints to the table. Then, the survey results are treated as the Ground Truth, setting a standard to measure how accurate the model is. The Ground Truth is then compared with the model's results using a confusion matrix, assisting in the thorough evaluation of the 'irony detection' process.

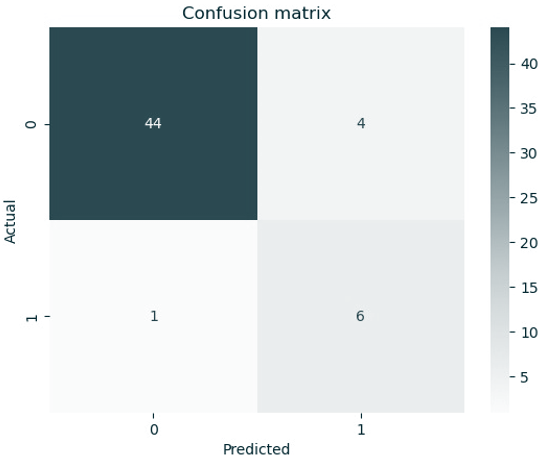


Illustration 2 – Irony Confusion Matrix

To evaluate the accuracy of "irony detection," a confusion matrix is utilized. This tool provides insight into the model's ability to predict irony. The matrix is divided into four sections:

**True Positive (6/7):** This section shows where the model accurately identified 6 out of 7 ironic tweets, in line with the survey results. This indicates the model's effectiveness in recognizing irony.

**False Positives (4):** Representing instances where the model incorrectly identified non-ironic tweets as ironic, this section comprises 4 instances. However, this is not a significant concern, as these are likely to be corrected by subsequent analysis by the LLM.

**True Negative (44):** The True Negative section highlights instances (44) where the model correctly identified non-ironic tweets. This signifies a robust performance in discerning non-ironic content.

**False Negatives (1/7):** In this section, the model missed identifying 1 ironic tweet out of 7. While a small error, it provides valuable insights for potential improvements in the detection process.

This breakdown within the confusion matrix offers a comprehensive understanding of the model's strengths and areas for enhancement in "irony detection."

***Process of Irony Evaluation***

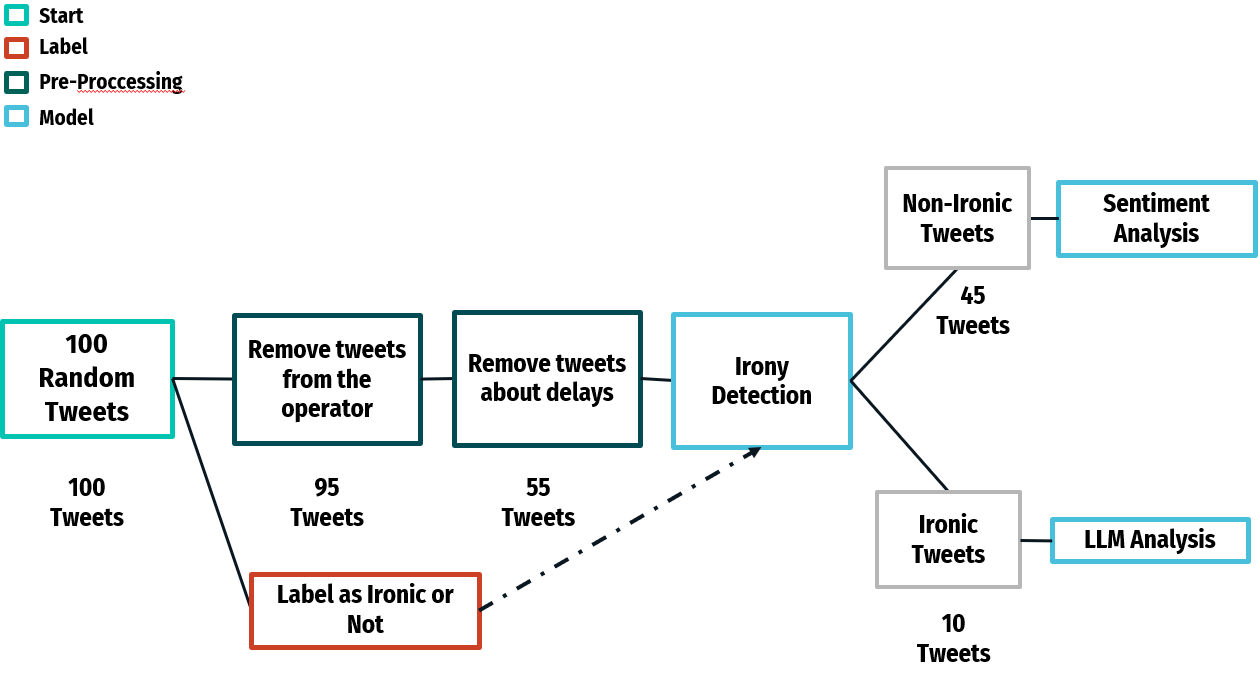


Illustration 3 – Process of Irony Evaluation

In the irony evaluation, 100 random tweets were chosen. From these, 5 tweets related to Thameslink operators and 40 concerning delays were excluded, leaving 55 relevant tweets. These tweets were then put through the irony detection process. The outcomes of this process were compared with insights from a survey, which served as the ground truth for the evaluation.

***Evaluation Scores of Irony***

In the summary of the irony evaluation, 5 evaluation scores were derived, with a particular emphasis on the key metric of Recall. Recall measures the ratio of true positives to the total of true positives and false negatives, assessing the model's capacity to identify relevant instances in the dataset. A high recall score, such as the achieved 85.7%, indicates the model's effectiveness in minimizing false negatives and accurately identifying most positive instances. In this case, with a focus on True Positives, the recall percentage suggests a low likelihood of the model missing ironic tweets.

**Precision (60%) :** The model is correct in predicting positive instances 60% of the time.

**Recall (85.7%) :** The model captures approximately 85.7% of all actual positive instances.

**Specificity (91.7%) :** The model accurately predicts negative instances about 91.7% of the time.

**Accuracy (90.9%) :** The model's predictions are correct around 90.9% of the time.

**F1 score (70.6%) :** Provides a balanced measure of both precision and recall, giving an overall indication of the model's performance.

**Evaluation of Sentiment Analysis**

***Defining Negative Sentiment***

Negative sentiment refers to a generally pessimistic, critical, and dissatisfied viewpoint.

***Significance of Identifying Negative Sentiment***

Understanding and identifying negative sentiment tweets is crucial for:

* **Enables Early Problem Detection:** It is crucial for quickly identifying potential issues. Early detection can lead to prompt intervention, which helps prevent small concerns from becoming larger problems.
* **Increases Customer Satisfaction:** It is essential in showing customers that their feedback is valued. By addressing negative comments effectively, it demonstrates a commitment to customer care, which can significantly enhance customer satisfaction.
* **Maintains a Positive Situation:** It is important for preventing negative experiences from worsening.

***Evaluation of Sentiment Analysis***

A two-stage method such as Irony Evaluation is used to evaluate how well "Sentiment Analysis Evaluation" works. The only difference is that the survey questions are Complaint & Non-Complaint instead of Ironic & Non-Ironic.

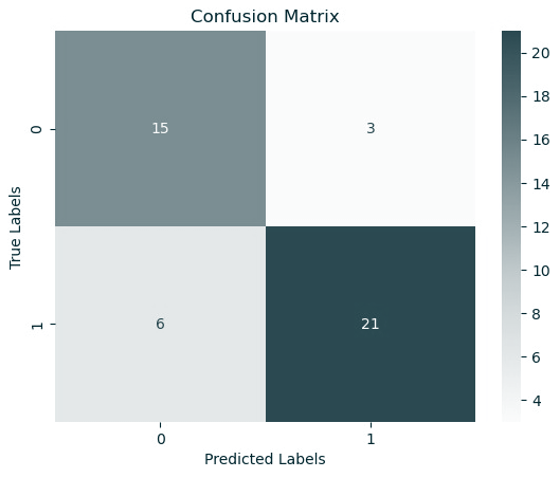


Illustration 4 – Confusion Matrix of Sentiment Analysis

To evaluate the accuracy of "Sentiment Analysis" a confusion matrix is utilized. This tool provides insight into the model's ability to predict negative sentiment. The matrix is divided into four sections:

**True Positive (21/30):** This part indicates the instances where the model accurately identified 21 out of 30 tweets as having negative sentiment, aligning with the survey findings. It highlights the model's capacity to correctly pinpoint negative sentiments.

**False Positives (3):** In this category, there are 3 cases where the model incorrectly labeled tweets as negative. While these are errors, they are generally not a major concern, as subsequent LLM analysis is likely to reclassify them as non-complaints.

**True Negative (15):** The True Negative section shows 15 instances where the model rightly identified tweets as positive or neutral. This reflects the model's strength in discerning tweets that do not carry negative sentiment.

**False Negatives (6):** The model missed 6 instances of negative sentiment. These omissions, while not numerous, are informative for potential enhancements to the model's detection capabilities.

***Process of Sentiment Analysis***

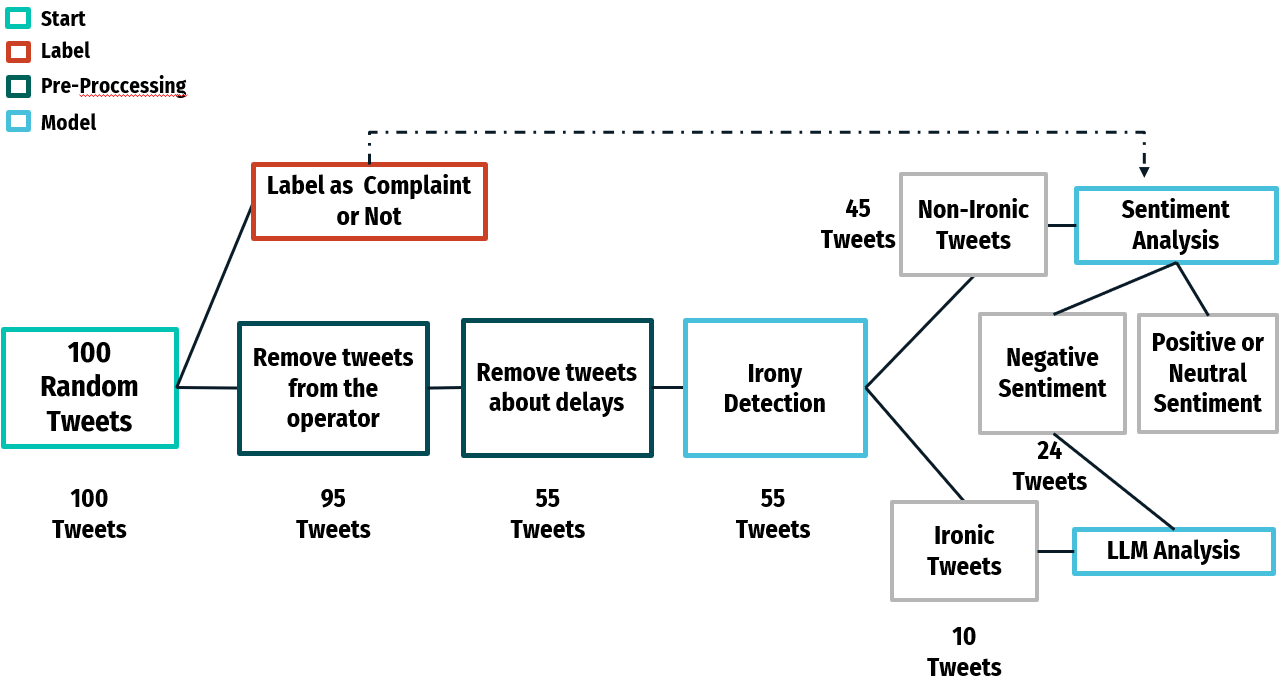


Illustration 5 – Process of Sentiment Analysis

In the evaluation process, 100 random tweets were initially gathered. Out of these, 5 tweets concerning Thameslink operators and 40 regarding delays were excluded, resulting in a total of 55 tweets deemed relevant for analysis. These selected tweets were then processed through irony detection. The output of this stage was 45 tweets that were classified as non-ironic and were subsequently analyzed for sentiment, being evaluated as negative, positive, or neutral. Tweets identified with negative sentiment were then directed into LLM Analysis.

***Evaluation Scores of Sentiment Analysis***

In the sentiment analysis evaluation, several scores were calculated, with each offering insights into the model's performance, particularly in detecting negative sentiment.

**Precision (71.4%)** : The model shows reliability in predicting positive sentiment, with 71.4% of its positive predictions being accurate.

**Recall (83.3%)** : The model successfully identified 83.3% of the actual negative tweets, as indicated in the survey, though it missed some positive tweets.

**Specificity (77.8%)** : The model's ability to recognize non-negative sentiments is reasonably good, with an accuracy of 77.8% in these predictions.

**Accuracy (80.0%)** : Overall, the model demonstrates an accuracy of 80.0%, taking into account both positive and negative predictions.

**F1 Score (76.9%)** : The F1 Score, which balances precision and recall, stands at 76.9%, indicating a solid overall performance of the model.

The results of the Sentiment Analysis Evaluation offer valuable insights. The Recall score, at 83.3%, indicates the model's efficiency in identifying negative tweets, based on the survey. This high percentage demonstrates the model's effectiveness in aligning with the ground truth in detecting negative sentiments.

**Evaluation of LLM Analysis**

***Purpose of LLM Analysis***

LLM Analysis is employed to thoroughly analyze passenger tweets, with a focus on efficiently categorizing and resolving train maintenance issues. A key part of this process involves interpreting irony within the tweets to ensure an accurate assessment of sentiment. This approach is vital for understanding the true nature of passenger feedback and addressing maintenance-related concerns.

***Evaluation of LLM Analysis***

The evaluation of LLM Analysis centers on testing its reliability. This is done by checking if the model provides consistent outputs when presented with identical inputs. Consistency in response is crucial as it indicates the model's stability and reliability in processing and interpreting information. By ensuring that similar inputs yield similar results, the reliability of the LLM Analysis in handling and categorizing tweets for maintenance issues can be confidently assessed.

***LLM Analysis Result Evaluation***

The evaluation indicates that LLM Analysis functions effectively, as demonstrated by:

**Consistent Formatting Results:** The analysis consistently delivers well-formatted outputs, ensuring reliable data interpretation.

**Effective Topic Categorization:** It adeptly categorizes topics, which is crucial for providing meaningful suggestions to the maintenance team. This categorization allows for the efficient distribution of tasks among relevant departments.

**Detailed Maintenance Suggestions:** Detailed and practical suggestions are provided, tailored to the needs of the maintenance teams, facilitating informed decision-making.

**Handling Irony:** The ability of the analysis to interpret and manage irony in tweets is a noteworthy feature, contributing to accurate sentiment analysis.

***Necessary Future Actions***

Looking ahead, certain actions are identified as essential for the continued success and improvement of the LLM Analysis:

**User Feedback Gathering:** Collecting feedback from users is critical to evaluate the real-world impact and effectiveness of the project outputs.

**Continuous Improvement:** There is a need for ongoing refinement of the model, incorporating new datasets, feedback, and corrections. This approach is vital to ensure the stability and enhancement of the model's capabilities.

**Evaluation of KPIs**

The evaluation of Key Performance Indicators (KPIs) in the project is structured under three main categories: KPIs, Impact of Project, and Resulting Effect.

**Operational Efficiency:** This KPI measures how the project streamlines the maintenance process. The resulting effect observed is faster problem resolution in trains, leading to an increase in overall operational efficiency.

**Cost of Maintenance:** By facilitating efficient issue resolution, the project helps in cost-saving. This efficiency translates into the company benefiting from reduced maintenance expenses.

**Equipment Reliability:** The project's swift problem-solving approach enhances the reliability of train equipment. As a result, trains become more reliable, providing safer transportation services.

**Safety Performance:** Timely maintenance ensures optimal safety. The resulting effect is a decrease in operational risks and an improvement in safety performance.

**Downtime Reduction:** The project's quick identification and resolution of issues positively affect train downtime. Consequently, train downtime is reduced, leading to increased customer satisfaction.

**Customer Satisfaction:** Improved service quality and prompt issue resolution indirectly influence customer satisfaction. This leads to passengers experiencing better service, which in turn results in higher customer satisfaction.

In summary,the Tweets2GPT project has been effective in detecting irony and negative sentiment within passenger tweets, contributing significantly to the maintenance process of Thameslink trains. The LLM analysis plays a pivotal role in this process, assisting the maintenance team by organizing similar tweets and offering detailed recommendations for action.

**Effectiveness in Detection***:* The project has demonstrated its capability in accurately identifying irony and negative sentiment in tweets, which is crucial for understanding and addressing passenger concerns related to train maintenance.

**Role of LLM Analysis:** LLM analysis aided the maintenance team by categorizing tweets with similar content and providing comprehensive suggestions, thus it streamlined the maintenance workflow.

**Need for Continuous Improvement:**To maintain and enhance the success of the project, there is a critical need for ongoing user feedback and continuous improvement. Incorporating new data and adapting to feedback are essential steps for the project's sustained effectiveness and relevance.