Introduction – Problem Statement

Hello, this is Steven from Thameslink. In times of ChatGPT I have the idea to use social media tweets to improve our maintenance. However, sarcasm influences the data quality in a negative way. Are you able to design and develop a system which identifies and interpret sarcastic texts/tweets ?

Pain Points -

1. Sarcasm affects Sentiment Analysis
2. Biased Data
3. Inaccuracy in Customer Satisfaction Analysis
4. And more....

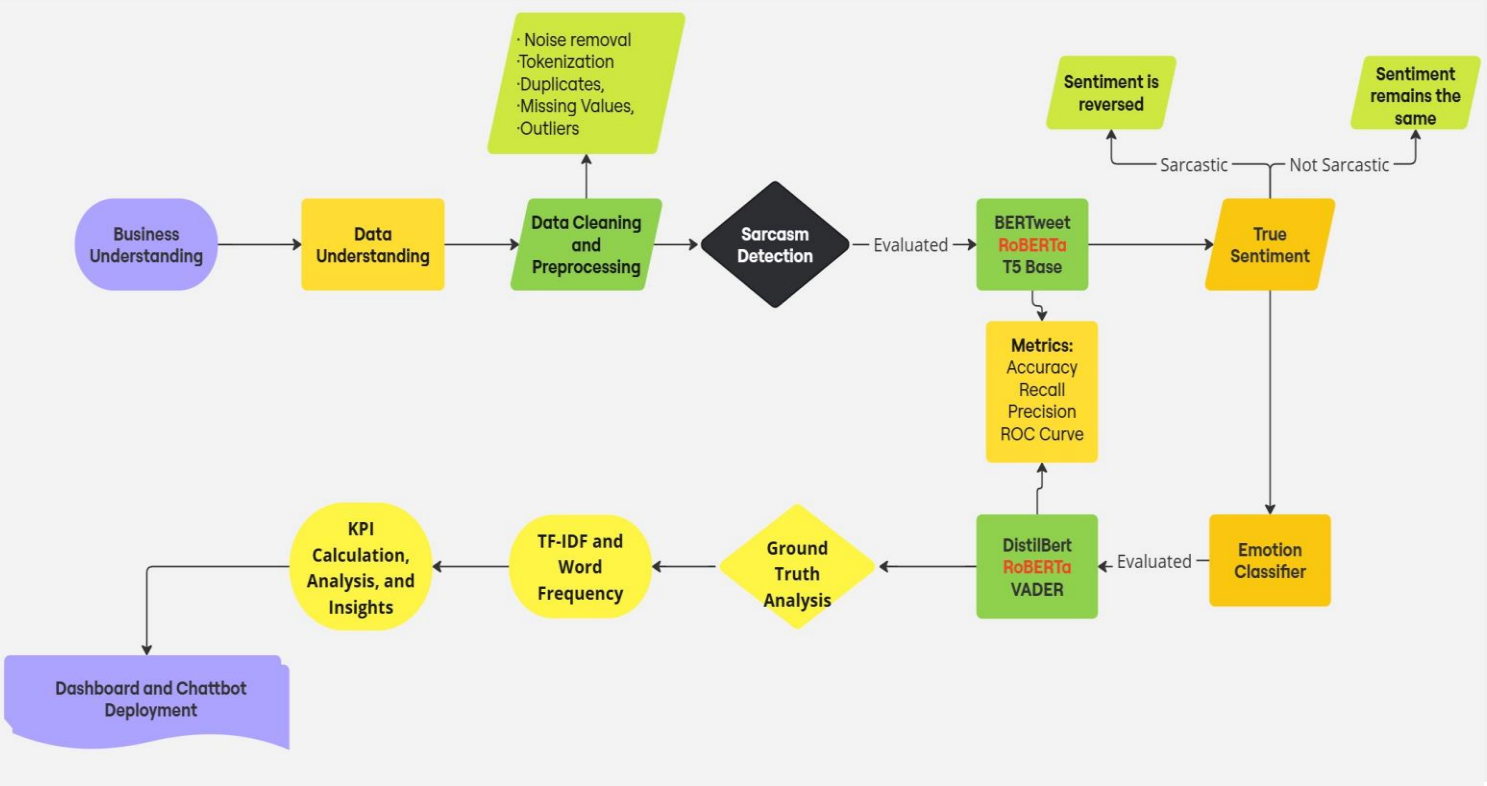
Tasks Contribution of Team Members

Topic	Dhruvi	Jui	Nishtha	Bhoomika	Abhishek	Naman
Business Understanding	17%	17%	17%	17%	17%	17%
Data Understanding	17%	17%	17%	17%	17%	17%
Data Preparation	17%	17%	17%	17%	17%	17%
Modeling	17%	17%	17%	17%	17%	17%
Evaluation	17%	17%	17%	17%	17%	17%
Deployment	17%	17%	17%	17%	17%	17%
Report Writing	17%	17%	17%	17%	17%	17%

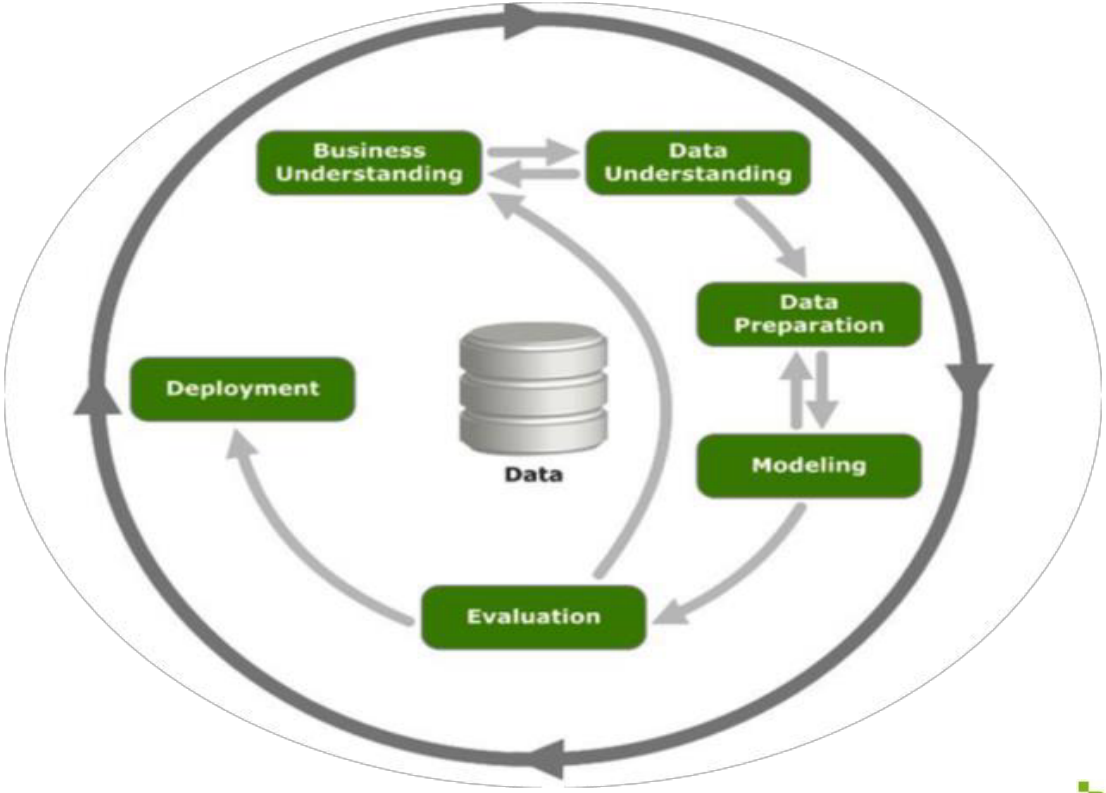
INDEX

Hochschule für Technik und Wirtschaft Berlin
University of Applied Sciences

1.1 Entire Project Methodology Flow



1.1 Cross-Industry Standard Process for Data Mining (CRISP DM)



1.2 Team Structure and Task Planning

Project Management Tool

We chose Asana as our project management tool. This not only allowed us the break down the project tasks per sprint, efectively and transparently, but also kept us informed what other team members were upto throughout the project. In the initial meeting, we would divide the tasks and later in a follow up meeting, give feedback to each other‘s results and close a task ticket, making execution quite swift.

Team Structure

We dicided team roles as per our previous professional experiences. However, while defining work structures, we made sure that all members get a chance to wear managerial and technical hats.

Task Planning and Execution

Tasks were defined during the sprint meetings and then as per expertise of team members, were divided among each other. We often worked in pairs too, so if one member is less experienced in a particular topic, they could learn better in collaboration with another member having sound expertise. This methodology of task distribution allowed every member to play around and develop confidence throughout all the steps in the project.

Collaboration Tools

We used a bunch of different tools for collaborating together-
Google collab and Jupyter notebooks for programming,
Microsoft Teams for meetings and file sharing,
Python as the programming language,
Power BI for dashboard design and development.

Board

htw. Hochschule für Technik und Wirtschaft Berlin
University of Applied Sciences

1.3 EPICS, user stories and acceptance criteria

CRISP-DM Phase	Epic	User Story	Acceptance Criteria
Business Understanding	Define the business problem and objectives.	**1.1: As a project manager**, I want to understand how sarcasm in tweets affects the business outcomes like vehicle availability.	A documented analysis that quantifies the potential impact of sarcasm on decision-making. A clear outline of specific business outcomes affected by sarcastic content.
	Clarify the requirements for detecting sarcasm to improve data quality.	**1.2: As a stakeholder**, I want to clarify the specific requirements for sarcasm detection in social media, to serve the business purpose, like improvement in cleanliness, delays, availability.	A requirements document listing key accuracy targets (e.g., 85% sarcasm detection accuracy). Specific scenarios and business needs for sarcasm detection are fully defined.
Data Understanding	Collect and explore relevant data.	**2.1: As a data engineer**, I want to collect a dataset of tweets relevant to the topic of interest, so that I can transform this for further analysis.	At least 10,000 tweets related to the topic are collected with relevant hashtags. Tweets are stored in a structured, accessible format (e.g., CSV or database).
	Explore and analyze the nature of sarcasm in the data.	**2.2: As a data scientist**, I want to explore the dataset for sarcasm-related patterns in the tweets, so that I can understand better.	An exploratory analysis report with key sarcasm indicators, sentiment distributions, and word clouds. A list of features or attributes indicative of sarcasm in the tweets.
Data Preparation	Clean and preprocess the data for modeling.	**3.1: As a data engineer**, I want to clean and preprocess the tweet data by removing duplicates and noise, so that they do not malign the dataset for training and further steps.	Dataset is free of duplicate tweets and irrelevant information. Preprocessed data is saved in a ready-to-use format for modeling (e.g., tokenized text). Like removing stop words, commas, period.
	Label the dataset for sarcasm and non-sarcasm tweets.	**3.2: As a data scientist**, I want to label tweets for sarcasm to prepare for model training, so that the data is proberly labelled for analysis using specific models.	Tweets are manually labeled for sarcasm and non-sarcasm.(In our case-labelled) A quality check is performed on the labels, achieving over 90% inter-annotator agreement.
Modeling	Develop and train sarcasm detection models.	**4.1: As a data scientist**, I want to build a baseline sarcasm detection model using NLP techniques. So that I can segregate the tweets based on the model and interpret their meaning, leading to improved services.	A baseline model is trained on the dataset. The model achieves a minimum accuracy of 70% on validation data.
	Tune and improve the model's performance.	**4.2: As a machine learning engineer**, I want to fine-tune the sarcasm model to improve accuracy, so that it can segregate the right tweets into categories.	The model is fine-tuned using hyperparameter optimization techniques (e.g., grid search, random search). The tuned model achieves at least 85% accuracy on validation data.
Evaluation	Evaluate different models and algorithms for sarcasm detection.	**4.3: As a data scientist**, I want to compare different NLP models (SVM, transformers, etc.) for sarcasm, so that I can choose which one brings the best results.	A report is generated comparing multiple models (e.g., SVM, transformers) on performance metrics like accuracy, precision, and recall. The best-performing model is identified.
	Test and validate the model's effectiveness.	**5.1: As a data scientist**, I want to evaluate the sarcasm detection model on unseen data to check accuracy, so that I can be sure which model is working.	Model performance is evaluated on a separate test dataset, with metrics calculated (accuracy, precision, recall, F1 score, confidence intervals). The model achieves set targets on unseen data.
	Assess how well the model solves the business problem.	**5.2: As a project manager**, I want to ensure the model meets business goals for improved decision-making, so that It can be deployed and scaled.	A report detailing the model's impact on business objectives is created, with comparisons to initial metrics. Stakeholder feedback confirms alignment with business needs.
Deployment	Deploy the final model into production.	**6.1: As a machine learning engineer**, I want to deploy the sarcasm detection model to monitor new tweets, so that it can help the business achieve its goals, like improved services.	The model is deployed in production and integrated with a live tweet streaming API.**(Concept Only)** Real-time sarcasm classification is functional and logs results in the database.
	Monitor model performance post-deployment.	**6.2: As a system admin**, I want to monitor the model's performance in production to detect drift or errors, so that I can make some tweaks and fine tune in case need be.	A monitoring dashboard is set up to track key metrics (accuracy, latency). Alerts are configured to notify admins when performance falls below set thresholds.**(Concept only)**

1.4 Product Vision

PROJECT CANVAS

Purpose/Goal:

Develop a sarcasm detection and interpretation system to improve vehicle availability insights for Thameslink by analyzing social media feedback.

Primary Stakeholders:

Data scientists
NLP researchers
Thameslink operations team
Persons writing the tweet

Secondary Stakeholders:

End users of sentiment analysis and fleet management dashboards
Thameslink customers (indirectly, through improved service)
Researchers studying sarcasm in communication

Scope:

Inclusions: Detection of sarcasm in various text forms (social media posts, reviews, comments, etc.), analysis of linguistic features, and interpretation models, Sentiment analysis.

Exclusions: Non-textual sarcasm detection (e.g., tone of voice), sarcasm in non-English languages, and real-time sarcasm detection.

Deliverables:

A trained sarcasm detection model capable of interpreting the meaning behind sarcastic tweets related to Thameslink's services.

A report or research paper detailing the methodology, challenges, performance, and applications.

Deployment Concept-A dashboard with KPIs, displaying insights derived from tweet analysis.

Resources:

People: Product Owner, Scrum master, data engineers, data scientists

Data: tweets related to Thameslink. Technology: NLP tools (like Hugging Face Transformers), machine learning frameworks (TensorFlow, PyTorch)

Budget: To cover data acquisition, computing power, software licenses, and hiring of specialists.

Timeline:

Business understanding – 2 weeks.

Data understanding and preparation- 2 weeks.

Model Iteration- 2 weeks.

Evaluation- 2 weeks.

Deployment- 2 weeks.

Customer challenge representation wrap-up and feedback- 1 week

Risks:

Data limitations: Sarcasm can be context-specific, making it difficult to get consistent labeling in datasets.

Model interpretability: While detecting sarcasm may be achievable, explaining why a text is sarcastic could be challenging.

Ambiguity: Sarcasm can be subtle and might not always be clearly identified, leading to false positives/negatives.

Success Criteria:

Accuracy: Achieve a high level of accuracy in sarcasm detection.

Interpretability: The system should provide the interpretation of the sarcastic tweet. The model should work well on diverse datasets and provide relevant insights about Vehicle Availability.

User feedback: Positive feedback from stakeholders indicating that insights are actionable and helpful in improving vehicle availability

1.5 KPIs

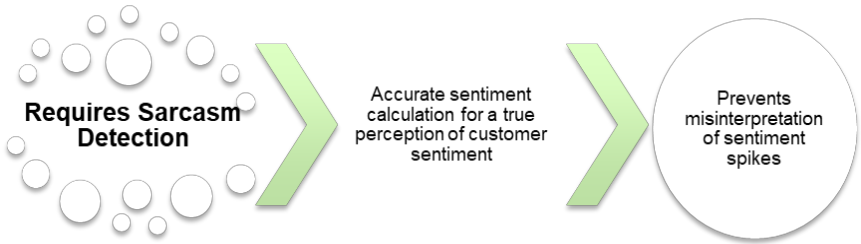

KPIs
Sentiment Analysis
Customer Satisfaction
Geospatial Analysis

1.5 KPIs (cont.)

Category	Metric	Description	Action Points
Category 1: Sentiment Analysis	**Tweet Sentiment Ratio**	• Calculate the ratio of positive, negative, and neutral sentiments based on the extracted sentiment labels.	• Sentiment Trends • Emotional Insights
	Sentiment Trends	• Tracks the volume of negative tweets over time (e.g., weekly, monthly, seasonal). • Reflects trend interpretation over time.	
	Average Sentiment Score	• Monitors overall customer perception over time by averaging sentiment scores.	
	Sentiment Change Rate	• Measures percentage change in sentiment compared to a previous period.	
	Comprehensive Sentiment Analysis	• Advanced NLP analysis to identify emotional tones such as anger or satisfaction. • Provides in-depth interpretation of emotions.	

1.5 KPIs (cont.)

Category	Metric	Description	Action Points
Category 2: Customer Satisfaction	**Tweet Volume by Topic**	• Counts occurrences of each topic in tweets (e.g., service, delays) over time.	• Operational Focus • Customer Priorities • Quality Control
	Customer Satisfaction Index	• Measures customer satisfaction based on positive feedback or sentiment.	
Category 3: Geospatial Analysis	**Location-based Complaint Analysis**	• Maps complaint hotspots using tweet locations to identify issue-prone areas. • Analyze geographic distribution of tweets based on latitude/longitude to identify hotspots of complaints.	• Regional Hotspots • Location-Based Insights • Targeted Improvements



1.6 Integration of Business Processes

The current (Manual) workflow

Customer service representatives - manually review & respond feedbacks (Identifying key issues like vehicle availability concerns).

Integration Point

Sarcasm Detection Tool
Captures and interprets customer emotions from feedback to enhance vehicle availability by helping teams identify potential problems faster.

Business process integration

Tweet data (customer feedback) --> Analyze data --> Model build --> Sarcasm identification --> Interpreting and Explaining the classification --> Improve vehicle availability by setting priority (From comments, check feelings like anger, irony)

- Customer Service Improvement: "Real-time insights prioritize urgent issues."
- Product Development: "Guides enhancements based on customer sentiments."
- Vehicle Availability: "Ultimately guides the organization in taking proactive adjustments to service schedules and management.

1.7 MoSCoW analysis

MUST HAVES

1. Sarcasm detection
2. Interpretation after sarcasm detection
3. Emotion Detection of sarcasm(Categorize them in the form of happy, sad, etc.)
4. interface for the product
5. Performance Monitoring and Accuracy Tuning
6, Data Preprocessing Pipeline
7. Language Standardization
8. Dashboard with action KPIs

SHOULD HAVE

1. User profiling based on recurring sentiment
2. Compalints on the basis of location
3. Top Complainants Identification
4. Tweet Sentiment Ratio
5. Keyword-Based Sentiment Analysis

COULD HAVE

1. Feedback on the comments
2. Anticipating Service Needs as per the sarcasm detection
3. Chat bot

WON'T HAVE

1. Voice Analysis of Customer Feedback Calls
2. Real-Time Sentiment Analysis
3. Sarcasm Detection in Visual Content (No Multi model)
4. Sarcasm Detection in Multiple langugaes

1.8 Risk

Risk Analysis

Data Risks

- **Poor Data Quality**: The model's performance can suffer if the training data is noisy, unbalanced, or inaccurate. Sarcasm detection heavily depends on nuanced contextual data, and poor-quality data can lead to unreliable predictions.
- **Privacy Non-Compliance**: Issues may arise if the data used for training or deployment violates privacy regulations or does not comply with legal standards like GDPR or CCPA.

Model Risks

- **Detection and Interpretation Analysis**: Challenges exist in ensuring the model properly identifies sarcasm, as sarcasm often depends on subtle context, tone, or prior knowledge, which the model may not fully capture.
- **Interpretation Inaccuracy**: Even when sarcasm is detected, the model may misinterpret the intent or misclassify genuine content as sarcastic, leading to errors.

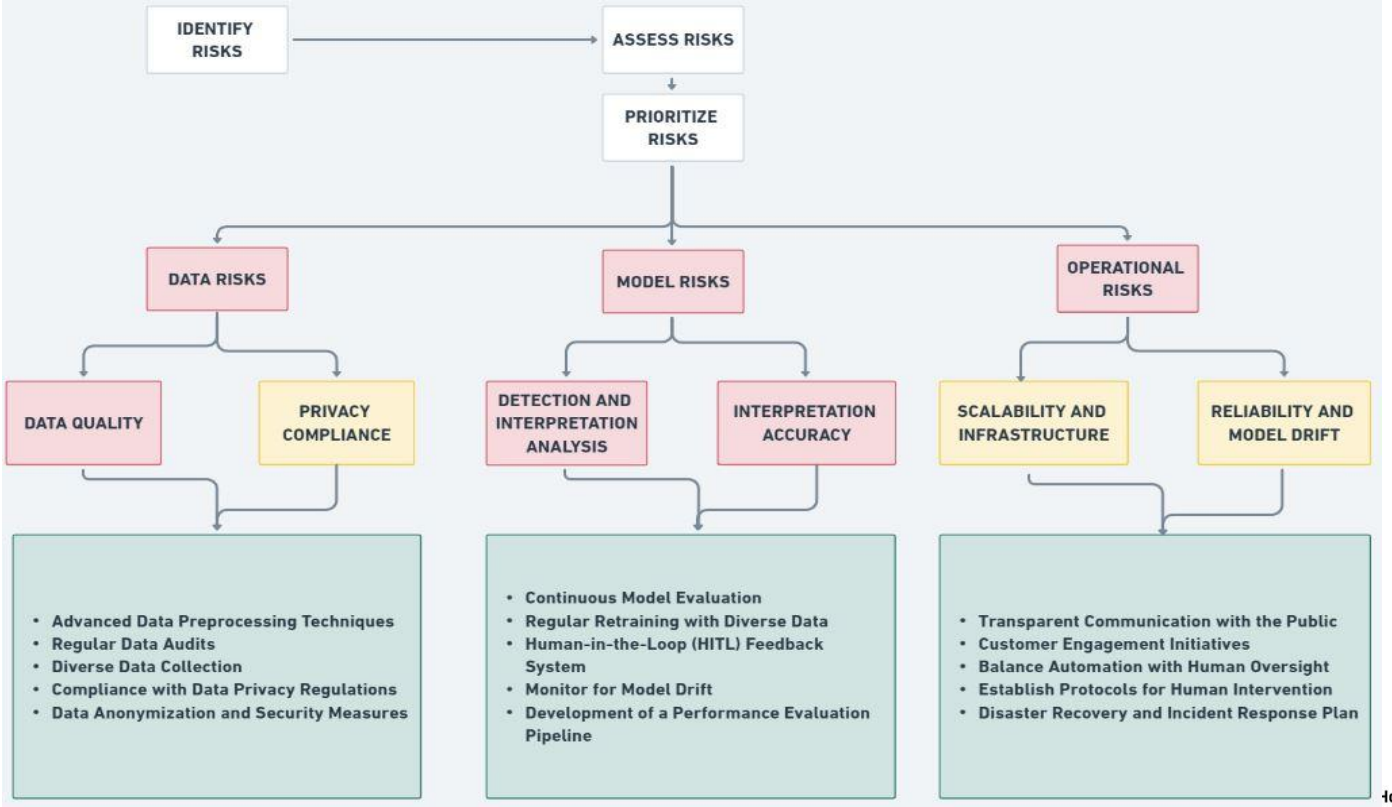
Operational Risks

- **Non-Scalability and Inadequate Infrastructure**: The system might struggle with large-scale deployment, especially if it requires extensive computational resources or real-time processing capabilities.
- **Non-Reliability and Model Drift**: Over time, the model may become less effective as language usage and sarcasm expressions evolve, requiring regular updates and retraining.

Risk Matrix

		LIKELYHOOD			
		RARE (1)	POSSIBLE (2)	LIKELY (3)	CERTAIN (4)
EFFECT	NEGLIGIBLE (1)	MINOR SOFTWARE ISSUES	VARIABILITY IN RESPONSE TIME		
	MODERATE (2)	PRIVACY INCOMPLIANCE	MODEL DRIFT	TECHNICAL FAILURES	
	SEVERE (3)		NEGATIVE PUBLIC PERCEPTION	DEPENDENCE ON AUTOMATION	RESULT INACCURACY ISSUES
	CATASTROPHIC (4)			POOR DATA QUALITY ISSUES	INADEQUATE INTERPRETATION ISSUES

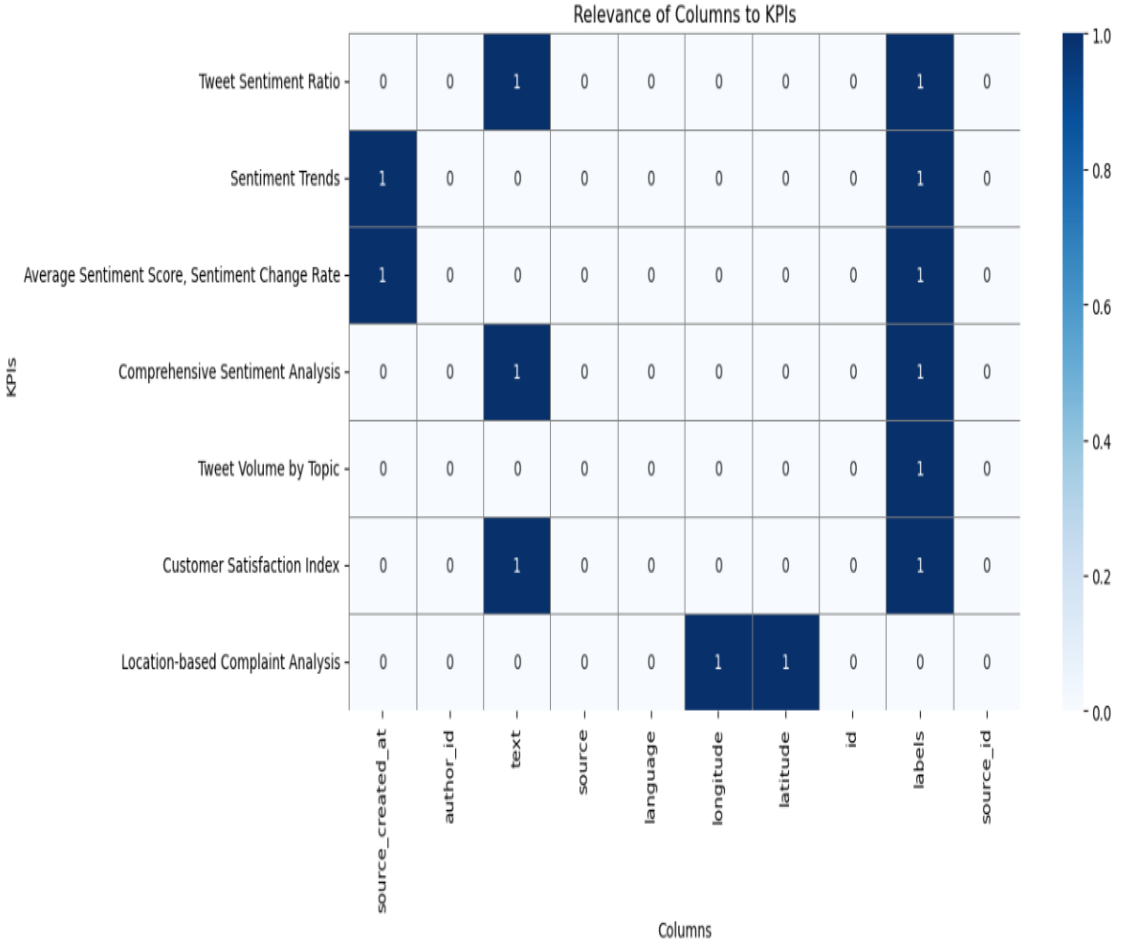
Mitigation


IDENTIFY RISKS
ASSESS RISKS
PRIORITIZE RISKS
DATA RISKS
MODEL RISKS
OPERATIONAL RISKS
DATA QUALITY
PRIVACY COMPLIANCE
DETECTION AND INTERPRETATION ANALYSIS
INTERPRETATION ACCURACY
SCALABILITY AND INFRASTRUCTURE
RELIABILITY AND MODEL DRIFT
• Advanced Data Preprocessing Techniques
• Regular Data Audits
• Diverse Data Collection
• Compliance with Data Privacy Regulations
• Data Anonymization and Security Measures
• Continuous Model Evaluation
• Regular Retraining with Diverse Data
• Human-in-the-Loop (HITL) Feedback System
• Monitor for Model Drift
• Development of a Performance Evaluation Pipeline
• Transparent Communication with the Public
• Customer Engagement Initiatives
• Balance Automation with Human Oversight
• Establish Protocols for Human Intervention
• Disaster Recovery and Incident Response Plan

2.1 Data dictionary

Column Name	Data type	Meaning	Useful for the KPIs
source_created_at	Integer (datetime)	The timestamp at which the tweet was made. (format - YYYY-MM-DD time)	Sentiment Trends Average Sentiment Score Sentiment Change Rate
author_id	Integer	The unique identification for each user.	This might use for user profiling
text	String	The username of the user and the tweet	Tweet Sentiment Ratio Comprehensive Sentiment Analysis Customer Satisfaction Index
source	String	The platform that was used to post the tweets.	Not useful in the defined KPIs
language	String	Language of the tweet from where the tweet was made.	Only for English GB data
Labels topic ground_truth sentiment	 String Boolean String	 Category of the tweet verified sarcasm If the tweet is positive, negative or neutral	Tweet Sentiment Ratio, Sentiment Trends Average Sentiment Score, Sentiment Change Rate, Comprehensive Sentiment Analysis, Tweet Volume by Topic, Customer Satisfaction Index
Source id	Integer	The username of the user	Not useful in the defined KPIs
longitude	Float	The east west location from where the tweet was made.	Location-based Complaint Analysis (Insufficient data)
latitude	Float	The north south position from where the tweet was sent.	Location-based Complaint Analysis (Insufficient data)

2.2 Relevance of columns to the KPIs



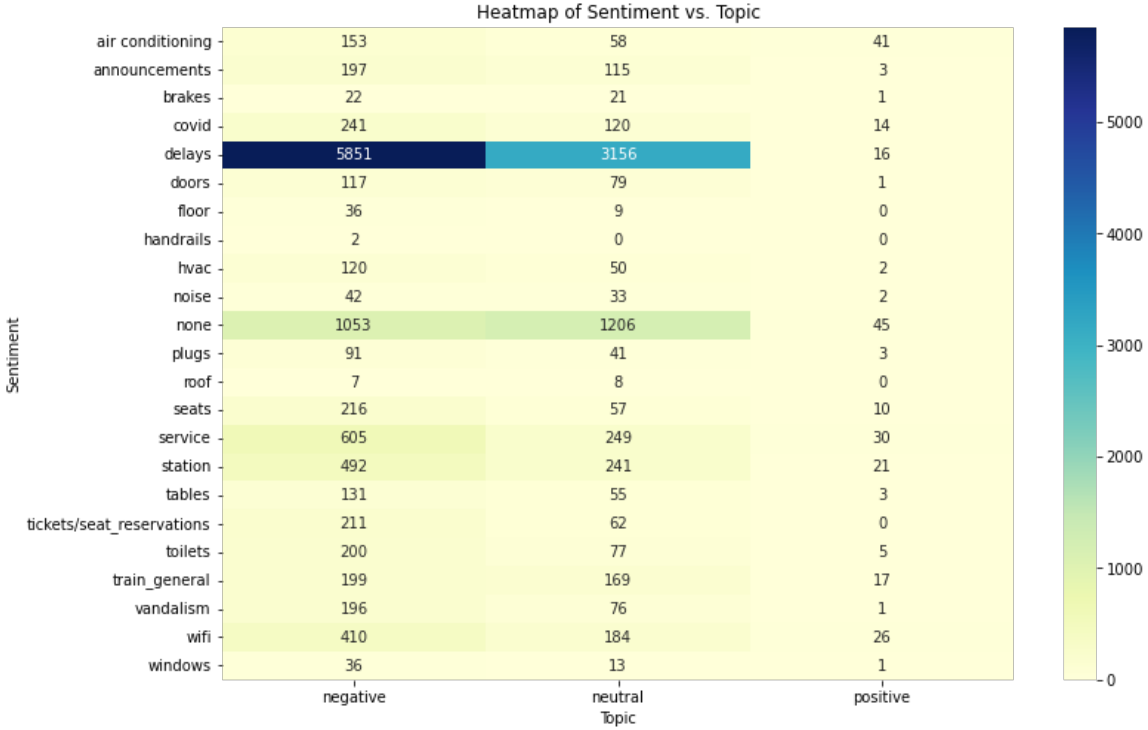
Understanding the relation between column to KPIs:

- Heatmap highlights the relevance of data columns to specific KPIs
- Provides clarity on which features are most impactful
- Useful in terms of dashboard creation for columns prioritization and irrelevant columns can be discarded.

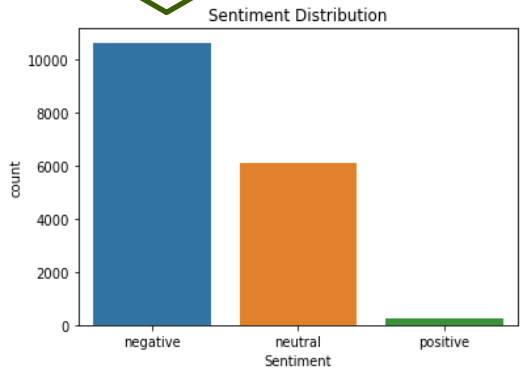
Impact: Columns marked as '0' will be excluded, ensuring streamlined data usage and improved KPI tracking.

2.3 EDA

1. Frequency Distribution



- This analysis helps to identify **frequently mentioned topics** tied to negative sentiments, allowing targeted improvements.
- Prioritization of topics like **delays, service, and station** can lead to more impactful interventions and improved customer satisfaction.



EDA script

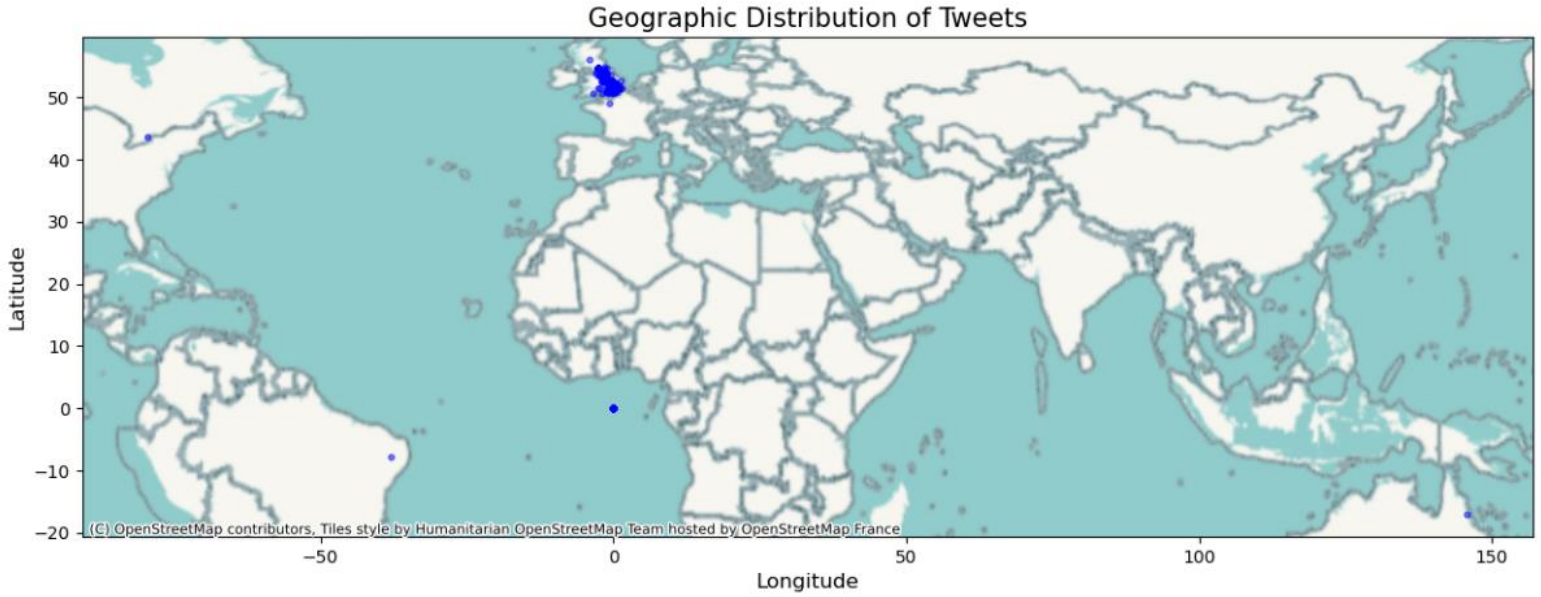
htw Hochschule für Technik und Wirtschaft Berlin
University of Applied Sciences

2.3 EDA (cont.)

Frequency Distribution of Countries:

Country	
United Kingdom	1391
Unknown	30
France	1
Australia	1
Brazil	1
Canada	1

- longitude and latitude helps identify **specific locations** where issues are reported, allowing companies to pinpoint problem areas.
- track **regional patterns** in customer behavior or issues, such as areas with consistently high negative sentiment.
- Allows location-specific action plans, like increasing vehicle availability at location.



2.3 EDA (cont.)

2. Duplicates

Finding	Solutions	Decision	Impact
1200 rows	1. Remove 2. Leave for Further Analysis	Duplicates were dropped.	1.29% data contains duplicate rows which can be removed for further analysis.

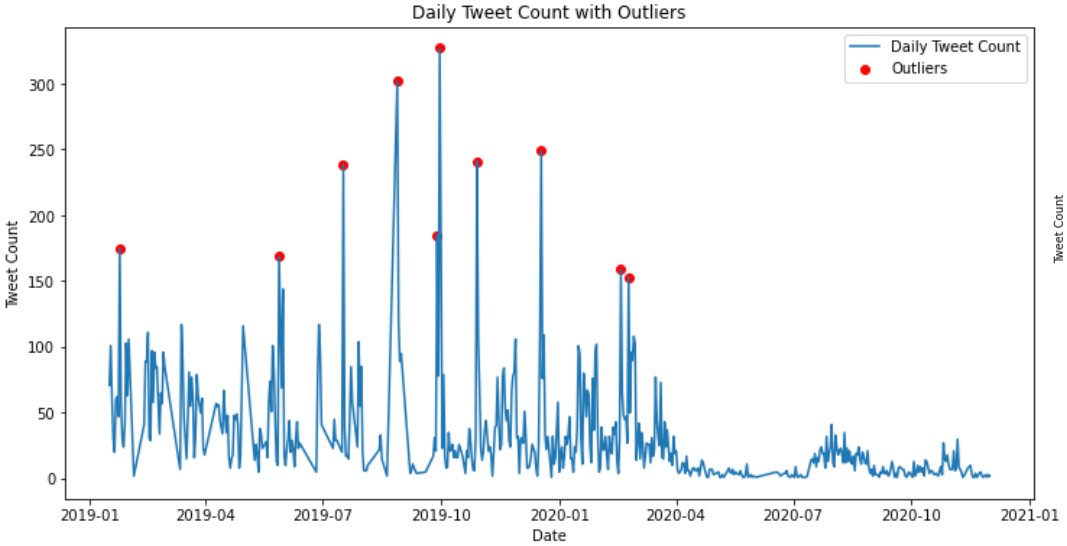
3. Missing Values

Finding	Solutions	Decision	Impact
Longitude-15524 Latitude - 15524	1. Remove 2. Leave for Further Analysis	Missing values were dropped.	91.6% data is missing hence cannot be considered for modelling.

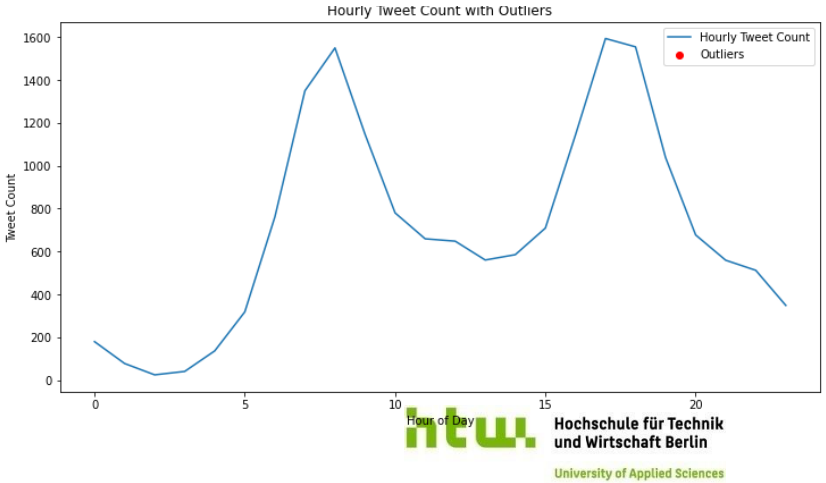
2.3 EDA (cont.)

4. Outlier Analysis of Tweet count on the basis of Time

Finding	Solutions	Decision	Impact
43 Outliers	1. Remove 2. Replace 3. Do nothing	3. Do Nothing	None



Peak hour analysis

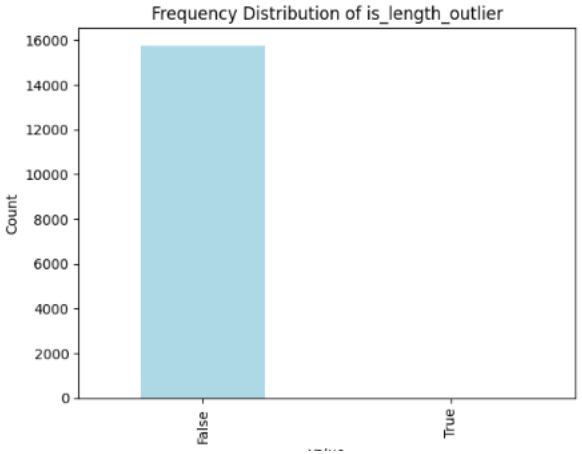


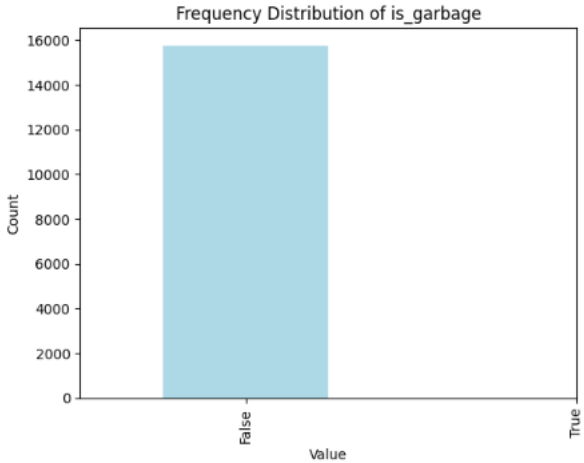
Hochschule für Technik und Wirtschaft Berlin
University of Applied Sciences

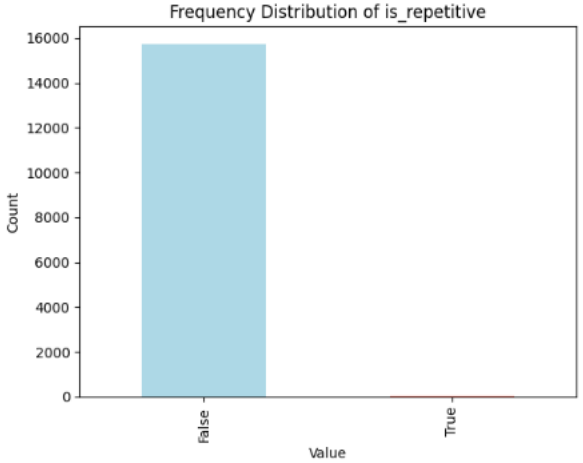
2.3 EDA (cont.)

5. Tweet Length Analysis

Finding	Solution	Decision	Impact
Outliers – none **Garbage Tweets-** none **Repetitive Tweets** - none	1. Keep Flagged Tweets 2. Discard Flagged tweets	No Action Required	None







2.4 Model Integration- Sarcasm Detector

The Sarcasm Detector is an advanced tool that leverages a pre-trained language model (like RoBERTa) to enhance the quality and interpretability of sentiment analysis data.

Raw Social Media Data → Actionable Business Intelligence

Key Functionalities

- **Sarcasm Score Calculation**: Assigns a numerical score (ranging from 0 to 1) to indicate the probability of sarcasm in each data entry.
- **Automated Labeling**: Classifies feedback as sarcastic (1) or non-sarcastic (0), enhancing the dataset's structure for analysis.
- **Text Analysis**: Processes complex language patterns to distinguish subtle sarcasm that could mislead traditional sentiment analysis tools.

Sarcasm_score	Sarcastic
0.345	0
0.807	1

Business Impact

Enhanced Sentiment Analysis: Provides more reliable input for tracking customer satisfaction KPIs, ensuring that negative sentiments disguised as sarcasm are appropriately flagged.

Data-Driven Decisions: Identifies areas where customer feedback has been misinterpreted, allowing for targeted improvements in service.

Reliability and Trust: Empowers businesses with dependable sentiment data for strategic decision-making.

Scalability and Efficiency: Provides a scalable solution that automatically processes and labels large volumes of feedback without manual intervention

Continuous Improvement: The sarcasm scores can be leveraged for continuous refinement of customer interaction models and strategies.

Hochschule für Technik und Wirtschaft Berlin
University of Applied Sciences

2.5 Data Cleaning & Preprocessing:
-Data Cleaning

Data Cleaning	Description	Findings	Possible Solutions	Decision	Impact
Missing Values	Identified and handled tweets with null or incomplete data to ensure data integrity.	Columns *Longitude & Latitude* contain 15,524 missing values.	1. Retain values 2. Drop Values	Retain values.	Geospatial Analysis KPI Cannot be represented as 91.6% datas missing.
Duplicate Tweets	Removed tweets with identical content or IDs to eliminate redundancy.	The dataset contains duplicates that could lead to biased results.	1. Keep all instances. 2. Retain only the first instance of each duplicate.	Retain only the first instance of each duplicate.	1200 rows were dropped, making tweet_id a Primary Key
Outlier Tweets	• Tweets with length over the threshold of +-3 std of mean length. • Outlier tweets on the basis of time analysis.	• None found. • 43 outliers found	1.Retain values 2.Drop Values	Retain outliers	--
Repetitive Tweets	Tweets with less than 3 unique words.	None found.	Drop Values	--	--
Garbage Tweets	Tweets containing only non-alphanumeric characters	None found.	Drop Values	--	--

2.6 Data Preprocessing

Data Preprocessing	Description	Purpose	Impact
Remove URLs	Strip out all hyperlinks (e.g., http://...).	Eliminate irrelevant content from tweets.	- Improves Model Accuracy - Reduces Dimensionality - Improves Speed and Efficiency - Improves Interpretability
Remove Mentions	Remove user mentions (e.g., @username).	Focus on the tweet content itself.	
Remove Hashtags	Remove hashtags (e.g., #example)	Avoid bias from hashtags in sentiment analysis.	
Normalize Whitespace	Replace multiple spaces with a single space.	Clean up the text for better readability.	
Convert to Lowercase.	Convert all text to lowercase	Standardize the text for consistent tokenization.	
Keep Stop words	Retain common words like "is," "not," "the," etc.	Keep stopwords to preserve subtle linguistic cues as they are crucial for sarcasm detection.	

	Columns	**Rows**
Raw Data	10	16,949
Cleaned & Preprocessed Data	18	15,749

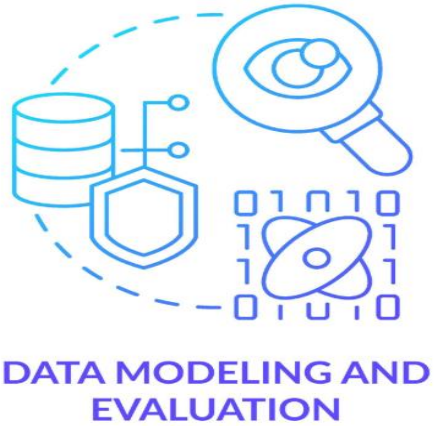
2.7 Modeling of Sarcasm Detector

Sarcasm Detector Model - Comparative Analysis

	Model 1 RoBERTa	Model 2 T 5 base	Model 3 BERTweet
Reason for Selection	Pre-trained on Twitter data, specifically designed for irony and sarcasm detection.	Fine-tuned for sarcasm detection on Twitter-specific datasets.	Trained on a combined sarcasm dataset, offering robustness across diverse sources.
Training Dataset Size and Composition	Trained on a large corpus of Twitter data, focusing on irony and sarcasm.	Fine-tuned on a specific subset of sarcastic tweets.	Trained on a combined dataset from multiple sources, enhancing generalizability.
Robustness to Noisy Data	Handles Twitter-specific noise well.	Sensitive to noise due to generative nature, requires data cleaning.	Robust against various noise types due to diverse training data.
Scalability	Optimized for Twitter data; may require retraining for other platforms.	Primarily suited for Twitter; scalability to other platforms needs evaluation.	Designed for cross-platform sarcasm detection, offering better scalability.
Hyperparameters	- Batch size: 16 - Learning rate: 5e-5 - Optimizer: AdamW	- Batch size: 8 - Learning rate: 3e-5 - Optimizer: AdamW	- Batch size: 16 - Learning rate: 2e-5 - Optimizer: AdamW
Challenges and Risks Identified	-Overfitting on small datasets. - Less robust on non-Twitter data.	- Over-dependence on token-level embeddings for sarcasm. - Slow	Requires significant computational resources for fine-tuning.
Real-World Usability	Easily integrates with Twitter-based systems; limited for other platforms.	Requires adaptation for real-time applications due to slower inference.	Versatile integration capabilities across platforms; higher resource needs.

3.1 Modeling and Evalution of the sarcasm Detection model

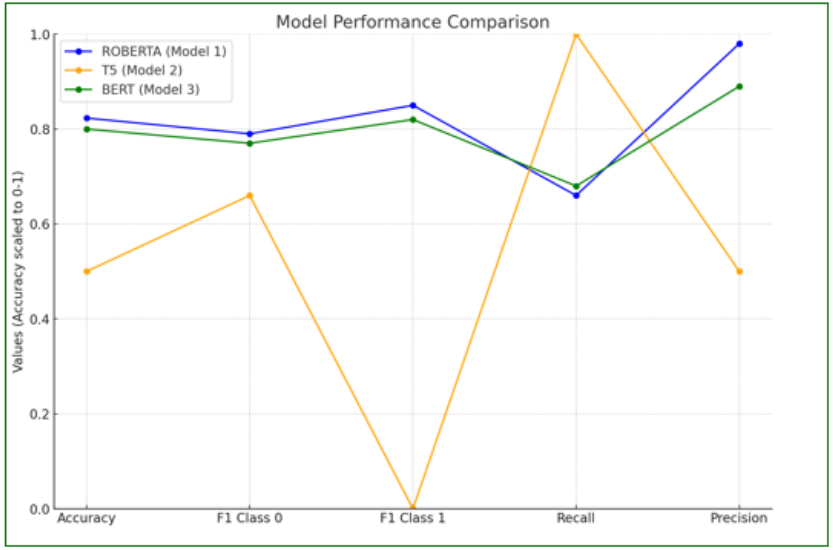
In the development of the sarcasm detection model, we utilized pretrained language models, including RoBERTa, BERT, and T5, to evaluate their performance. Among these models, RoBERTa achieved the highest accuracy of 82%, making it the most effective choice for sarcasm detection. Consequently, RoBERTa was selected for the subsequent interpretation and analysis processes, ensuring a focus on the model with the best performance.

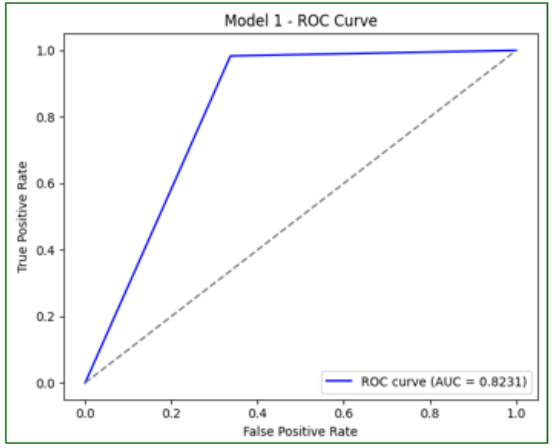


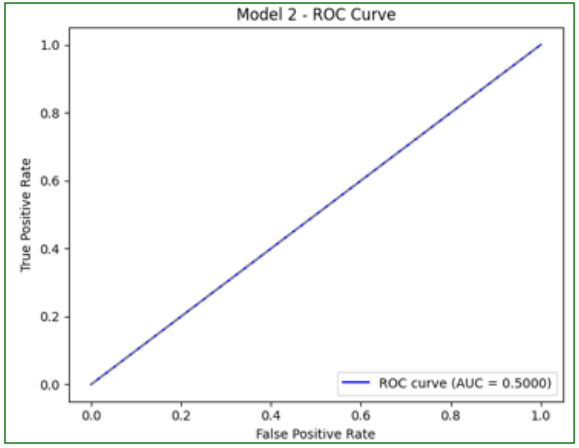
Script

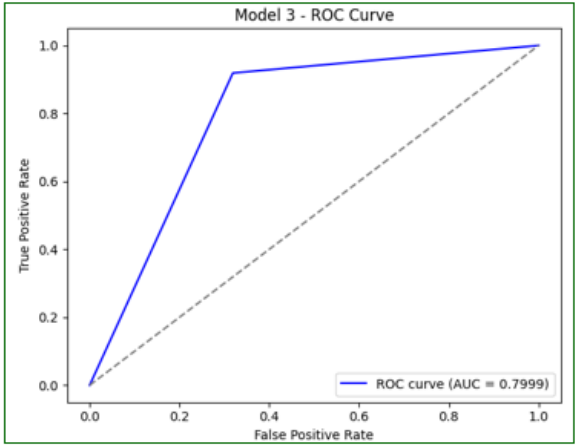
3.2 Model Performance

Key Metrics	Model 1 ROBERTA	Model 2 T5	Model 3 BERT
Accuracy	0.8231	0.50	0.80
F 1 score (class 0, class 1)	0.79 , 0.85	0.66 , 0.00	0.77 , 0.82
Recall	0.66	1	0.68
Precision	0.98	0.50	0.89









Model Performance Comparison

- ROBERTA performs well across all metrics.
- T5 gives poor performance, especially for class 1, which is evident in its F1 score dropping to 0 for that class.
- BERT - Good overall performance, but slightly lower recall compared to ROBERTA.

ROC Curves

- ROBERTA -> ROC AUC = 0.8231, indicating strong classification performance.
- T5 -> ROC AUC = 0.5000, equivalent to random guessing, confirming T5's poor performance.
- BERT ->ROC AUC = 0.7999, showing good performance, close to ROBERTA's.

Key Takeaways

- ROBERTA -> Best performer overall with high accuracy, balanced F1 scores, good recall, and strong ROC AUC.
- T5 -> Performs poorly across all metrics, especially in classifying class 1, evident from its zero F1 score for that class and ROC AUC of 0.5.
- BERT -> Performs well but slightly below ROBERTA in accuracy, recall, and AUC.

ROBERTA is the most effective model for this task, providing consistent and high-quality performance. BERT is a close second, while T5 fails to perform effectively.

htw. Hochschule für Technik und Wirtschaft Berlin
University of Applied Sciences

3.3 Best and worst Performing Models

Best Performing Sarcasm Detection Model
- RoBERTa

The model is **pre-trained specifically on Twitter data** for the task of irony detection. It has been exposed to the nuances of Twitter language, including slang, abbreviations, hashtag and informal expressions. This **domain-specific** training enables the model to better understand and detect sarcasm within the unique linguistic context of tweets.

RoBERTa's **architecture and training dataset** align closely with our problem and requirements, giving it a significant edge over more generalized models like T5 or BERTweet.

Simplicity in Task Execution, Robustness to Twitter-Specific Noise, Computational Efficiency. RoBERTa processes tweets faster while maintaining **high accuracy**, making it suitable for applications such as monitoring Twitter for customer feedback.

Worst Performing Sarcasm Detection Model
– T 5

T5 is designed to handle a wide range of tasks by reframing them as text generation problems. Sarcasm detection relies more on accurate classification rather than generative capabilities.

T5 processes require more computational resources compared leading to longer processing time. This can hinder usability, especially with large datasets.

T5's architecture is designed for multi-task learning and handling complex, multi-faceted tasks. For sarcasm detection—a relatively narrow classification problem—this added complexity is unnecessary.

Model selected for Sarcasm Detection: RoBERTa

Thameslink's needs to detect sarcastic tweets that distort customer satisfaction data.

RoBERTa's design directly addresses these pain points by focusing on sarcasm detection within this domain.

RoBERTa provides reliable and interpretable results that align with Thameslink's goal

3.4 Impact of Dataset Quality on Model Performance

Challenges Faced

Dataset Quality Issues: Duplicate entries, mislabeled samples, and imbalanced classes in the Kaggle dataset led to noisy, biased data.

Poor Performance: Low precision, recall, and F1-scores persisted, as shown by the confusion matrix, ROC, and classification report.

Frustration Point: Even after data cleaning and preprocessing, the model struggled to perform, raising doubts about the value of the pretrained model.

Time Cost: Extra effort was spent cleaning the dataset without significant improvement in results.

Applied Solution

Recognized the need for a high-quality, well curated dataset.

Downloaded a cleaner, balanced dataset with high-quality labeling for sarcasm detection.

Re-evaluated the same pretrained model on the new dataset.

Dataset 2 demonstrated a marked improvement in model performance, validating the impact of data quality.

Key Learnings

Quality Over Quantity: A smaller, cleaner dataset can significantly outperform a larger, noisy dataset.

Garbage In Garbage Out: The effectiveness of *pretrained* models like RoBERTa relies heavily on the quality of the data they're fine-tuned on.

Data Cleaning Has Limits: While cleaning improves data to some extent, poorly labeled or imbalanced data can still undermine performance.
Using high-quality datasets is critical for actionable insights.

Poor-quality data can waste resources and produce unreliable results.

RoBERTA with dataset 1

Classification Report:

	precision	recall	f1-score	support
Not Sarcastic	0.83	0.72	0.77	3934
Sarcastic	0.31	0.47	0.38	1066
accuracy			0.67	5000
macro avg	0.57	0.60	0.57	5000
weighted avg	0.72	0.67	0.69	5000

AUC Score: 0.5955

RoBERTA with dataset 2

Classification Report:

	precision	recall	f1-score	support
Not Sarcastic	0.98	0.66	0.79	884
Sarcastic	0.75	0.98	0.85	900
accuracy			0.82	1784
macro avg	0.86	0.82	0.82	1784
weighted avg	0.86	0.82	0.82	1784

AUC Score: 0.8231

3.5 TF-IDF ----> How often a word appears in a tweet and in a dataset

Decreasing the manual work	• Automated topic identification • Actionable insights
Further evaluation of emotions	• Emotion categorization • Customer prioritization
Improving vehicle availabilty	• Real time issue detection • Optimized resource allocation

Role of TF-IDF in Sarcastic and Normal Tweets

Extracting Key Themes

- Identify important words in sarcastic tweets (e.g., 'sure,' 'obviously')
- Highlight contrasting patterns in normal tweets.

Applications

- Understand thematic differences between sarcastic and normal tweets.
- Use top TF-IDF words to refine emotional insights and dashboards

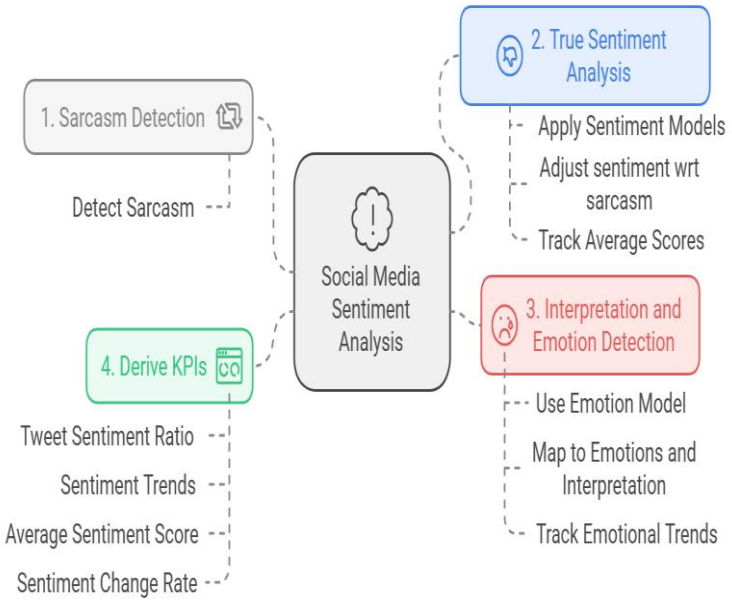
Word Cloud for Sarcastic Tweets

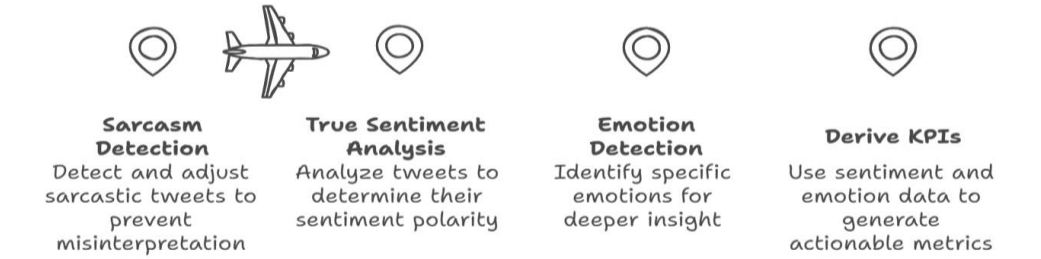
Word Cloud for Non-Sarcastic Tweets

TF-IDF Script

htw Hochschule für Technik und Wirtschaft Berlin
University of Applied Sciences

3.6 Emotion Detection

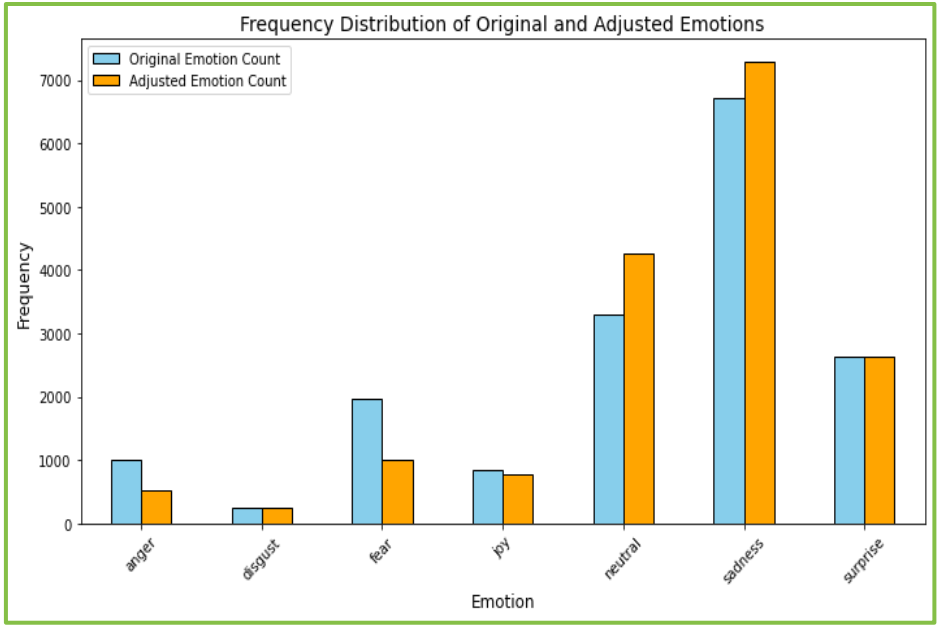




df

ɔpic	Topic ID	...	cleaned_text	Tokenized Text	sarcasm_score	sarcastic	VADER_neg	VADER_neu	VADER_pos	VADER_compound	Adjusted_Emotion	Original_Emotion
vice	8655353b-cc71-1e89-95d5-1c4e6bf200e1	...	the thameslink core between london st pancras ...	[the, thameslink, core, between, london, st, p...	0.340264	0	0.000	1.000	0.000	0.0000	sadness	sadness
days	8aee5c86-6e93-24d4-6d9d-d8a9f2f19cef	...	loving the complaint about people having to wa...	[loving, the, complaint, about, people, having...	0.941960	1	0.096	0.614	0.289	0.6597	sadness	joy

3.6 Emotion detection



Potential Steps-

- Evaluating sentiment models
- Updating code to just emotions- anger, fear, neutral, joy, and sadness

- **Fear**- "if you have a car park permit at a great northern station you may park at a thameslink station instead.due to the severity of the disruption, **we advise heavily that you do not travel**.if you do, **be prepared for cancellations, alternative routes and extended journey times.**"

- **Sadness**- "@TLRailUK The delay repay is asking for a required field when everything has been completed. **Can you process this please as the form doesn't seem to be working correctly**"

- **Anger**- "Today's train fun: woman sits on the seat in front of me and takes her mask off.In true british fashion, **I passive aggressively look her dead in the eye, stand up and move seats.C'mon** @TLRailUK, sort this out..."

- **Disgust**- "@BTPCambs @TLRailUK Take **more than a mask to persuade me to ride on your stinking trains**"

- **Surprise**- "#Thameslink & #govia is such a joke, **travel disruptions because of rain? really?**"

- **Joy**- "#Thameslink **congratulations** on the Cambridge to Brighton experience, **loving** the pared down no tables feel, the buttock-toning Soviet comfort of the seats and the eco-friendly lack of sockets! #puttingpassangersfirst"

3.7 Best and Worst Examples of Summarization Model

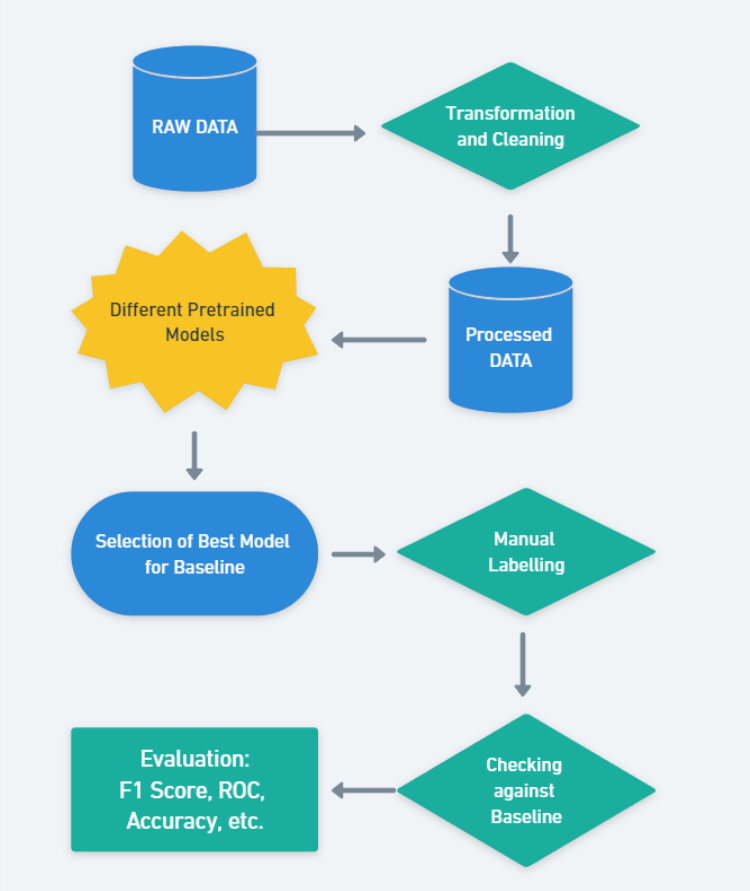
Actual	Output	Remarks
Hi @TLRailUK at Blackfrairs, tell your drivers on plat 2 to hold the doors open long enough, so passengers can leave a train terminating at plat 3, and run 20metres as soon as the doors open, to get on a connecting train. Shoddy service you run -I'll miss my connection now.#fail	i'm sorry, but i'm not sure if i'll be able to get on a connecting train	✓
@TLRailUK @SW_Help @SouthernRailUK @TfL Yyyiouuu FUCKING CUNTS.Just once it'd be nice to know my cunting train isn't cancelled for some cunting reason. You're the reason Britain is a third world country when it comes to trains. Piss poor service, expensive as fuck. GET FUCKKKKED!	Yyyiouuu yyyiouuu	✗
How do you get to Gatwick or Brighton tonight? #EastCroydon am baffled "Gatwick Express, Southern and Thameslink customers are strongly advised NOT TO TRAVEL via East Croydon. Please delay your journey until later, or use an alternative route" What's the alternative?	Gatwick Express, Southern and Thameslink customers are strongly advised NOT TO TRAVEL via East Croydon. Please delay your journey until later, or use an alternative route	✓
#TLRailUK morning train delayed and overcrowded. Evening train delayed. Morons at #thameslink that drive the trains always put in the final hard brake to throw you down the carriage. Employer and employee on harmony - BOTH INCOMPETENT.	The train is delayed and overcrowded.	✓
(Thameslink Update) 08:07 St Albans City to Sutton Surrey due 09:44 - 08:07 St Albans City to Sutton Su	The delay is being delayed at Tulse Hill.	✗
@Ellen30668723 @TLRailUK Hey, how much extra would you pay for the table?	i would pay a total of Â£20	✗

Discarded the idea of moving ahead with summarization model

Gave only active passive translations/ same output for most of the cases

Models are trained on the basis of standard language – could not generalise well the tweet language

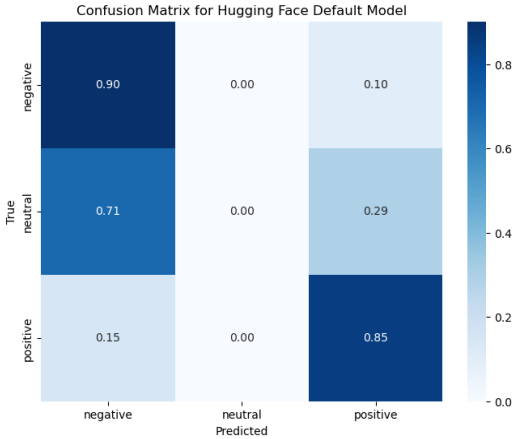
4.1 Evaluation Methodology Basic Plan

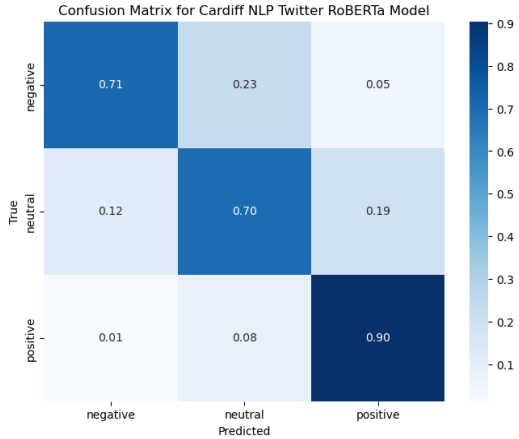


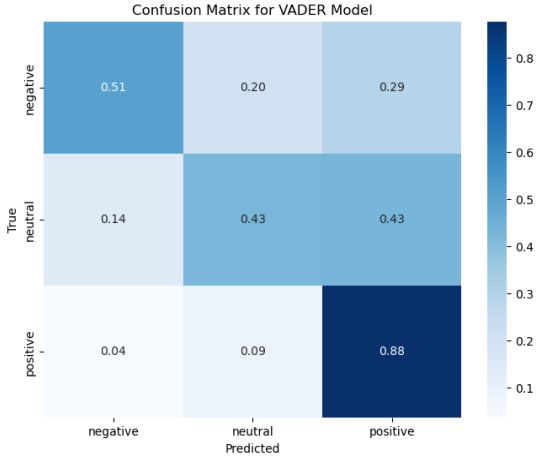
4.2 Sentiment Detection- Qualitative

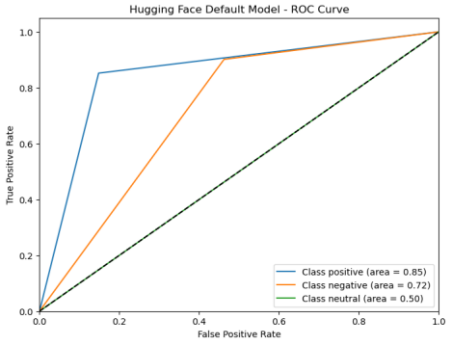
Aspect	distilbert-base-uncased-finetuned-sst-2-english	VADER Sentiment Analyzer	Cardiff NLP Twitter RoBERTa
Suitability for Social Media	Moderate **(not specifically trained on tweets)**	High (optimized for slang and emojis)	Very High **(specifically trained on tweets)**
Ease of Use	Requires pre-trained model setup (medium complexity)	Very Easy (out-of-the-box functionality)	Requires pre-trained model setup (medium complexity)
Speed	Moderate	Very Fast	Moderate
Handling of Sarcasm	**Limited**	**Poor**	**Good (trained on sarcasm in tweets)**
Output Categories	Positive, Negative **(No neutral)**	Positive, Negative, Neutral	**Positive, Negative, Neutral**
Advantages	- Pre-trained on high-quality datasets	- Lightweight and fast	- Specifically trained for tweets
	- High accuracy	- Simple to use	**- Handles informal text**
	- Handles longer text well	**- Interprets punctuation like "!" effectively**	- Better detection of sarcasm
Disadvantages	- Requires setup and infrastructure	- Rule-based, so it misses nuance	- **Requires GPU for faster inference**
	- Struggles with sarcasm and slang	- **Struggles with complex sentences and sarcasm**	- Setup is more complex
	- Not tailored for tweets	- Limited to short text	- Can overfit to Twitter-specific language
Best Use Case	General sentiment analysis for formal documents or reviews	Quick analysis of tweets, reviews, or comments	**Sentiment analysis for Twitter and social media**

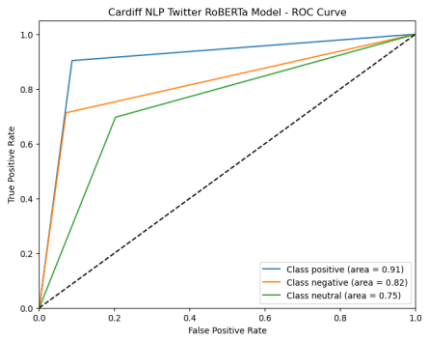
4.3 Sentiment Models Evaluation

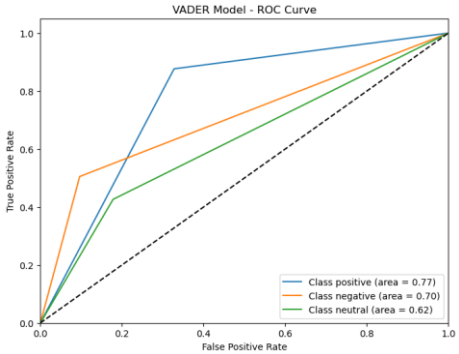




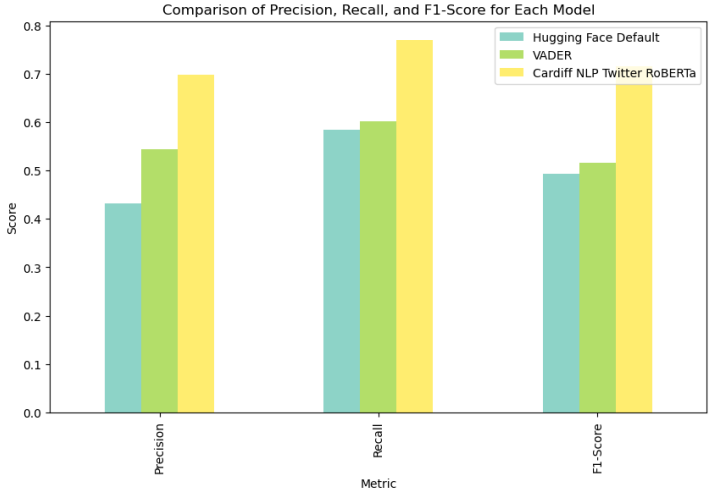


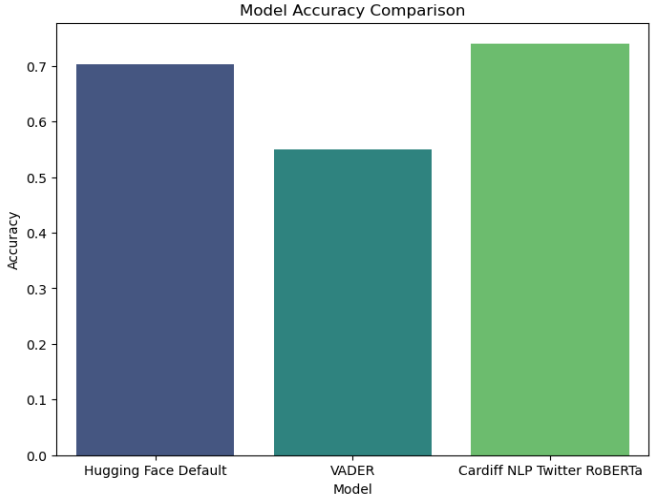






4.4 Sentiment Models Evaluation





Tweet ID	Sarcasm	Predicted_sentiment	Adjusted_sentiment
Swerving two rotten banana skins and a ghastly chewing gum seat, I cleared the coffee cups and moved in to my new home. At Â£40 an hour @TLRailUK was kind enough to bypass hygiene and leave detritus on the floor to play with as itâ€™s always delayed. #CoronaVirusUpdates https://t.co/L8Q1kRt8SJ	1	Negative	Positive

4.5 Sentiment Classification Evaluation- Quantitative

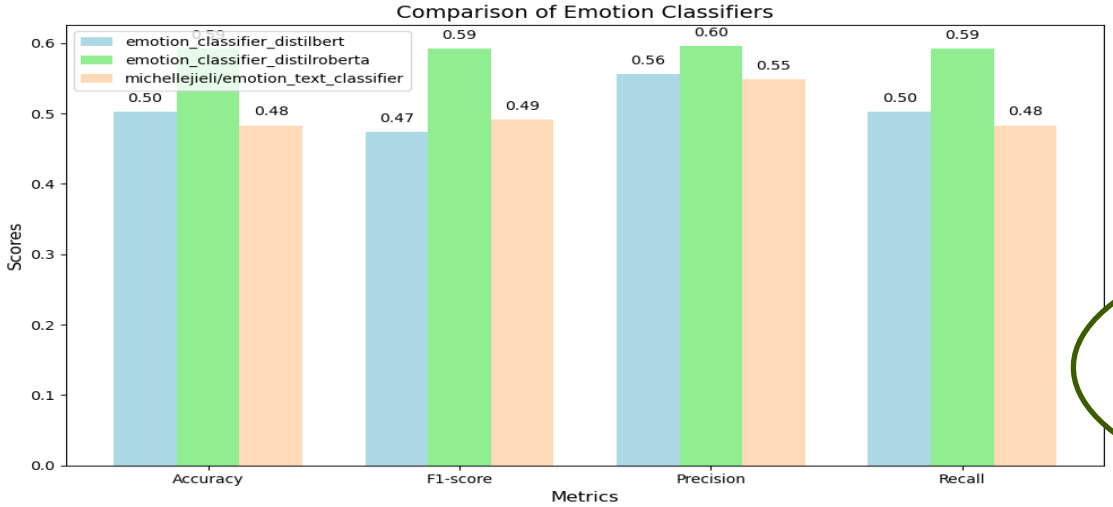
Category	Cardiff NLP Twitter RoBERTa	Hugging Face Default Model	VADER
Model Performance	- Best overall performance.	- Decent overall performance but fails for neutral sentiment (F1 = 0).	- Good for negative sentiment but struggles with neutral and positive sentiments.
	- High F1 scores: Negative (0.81), Neutral (0.57), Positive (0.77).	- F1 scores: Negative (0.83), Neutral (0.00), Positive (0.65).	- F1 scores: Negative (0.65), Neutral (0.41), Positive (0.49).
	- **High Accuracy: 74%.**	**- Accuracy: 70%.**	- **Lowest Accuracy: 55%.**
	- **Balanced classification for all classes**.	**- Struggles to detect neutral sentiment accurately.**	- **Imbalanced classification across sentiment categories**.
ROC Curves	- ROC AUC: Positive (0.91), Negative (0.82), Neutral (0.75).	- ROC AUC: Positive (0.85), Negative (0.72), Neutral (0.50).	- ROC AUC: Positive (0.77), Negative (0.70), Neutral (0.62).
	- **Best ROC AUC across all classes**, indicating robust classification performance.	- **Average ROC AUC performance**, with poor results for neutral sentiment.	- **Poor ROC AUC** for neutral and positive classes.
Key Takeaways	- Best performer overall, with strong accuracy, recall, F1 scores, and ROC AUC values.	- Performs well for negative and positive sentiments but fails in detecting neutral sentiments effectively.	- Effective only for negative sentiment detection but poor for neutral and positive sentiments.

Please find the script here: [Sentiment Evaluation Script]

4.6 Emotion classifier- Qualitative

Aspect	j-hartmann/emotion-english-distilroberta-base	bhadresh-savani/distilbert-base-uncased-emotion	michellejieli/emotion_text_classifier
Emotion Categories	**Joy, Anger, Sadness, Fear, Surprise, Love, Disgust**	Joy, Anger, Sadness, Surprise**, Love**	Joy, Sadness, Anger, Fear, Surprise, Love
Suitability for Social Media	**High (trained on emotion-rich datasets)**	Moderate (more general-purpose)	High (works well across formal and informal text)
Ease of Use	Medium (requires setting up Hugging Face model)	Easy to use with Hugging Face Transformers	Medium (pre-trained but requires fine-tuning for specific tasks)
Speed	Moderate	High (faster due to distilbert architecture)	Moderate
Handling of Nuance	**Captures overlapping emotions (e.g., anger + sadness) and complex, layered sentiments – Excellent**	Moderate (**struggles** slightly with nuanced emotions)	Good (handles nuanced emotions better than simpler models)
Advantages	- **Fine-tuned for rich emotion detection**	- Lightweight and fast	- Strong in multi-domain emotion classification
	- Captures nuanced emotions	- Simple setup	- Handles a variety of input styles
	- **Works well with short texts (e.g., tweets**)	- Balanced performance	
Disadvantages	- Moderate setup complexity	**- Limited emotion categories**	**- Medium speed, fails with noisy data**
	- Requires significant compute resources for batch inference	- **Struggles with highly nuanced or overlapping emotions**	- Requires additional fine-tuning for domain-specific datasets
Best Use Case	**Emotion classification in social media** or conversational datasets	General emotion classification for reviews or comments	Multi-purpose emotion analysis across diverse datasets
Output Format	Probabilities for each emotion (e.g., joy: 0.78, sadness: 0.12)	Probabilities for each emotion (similar structure)	Emotion probabilities and label with high confidence

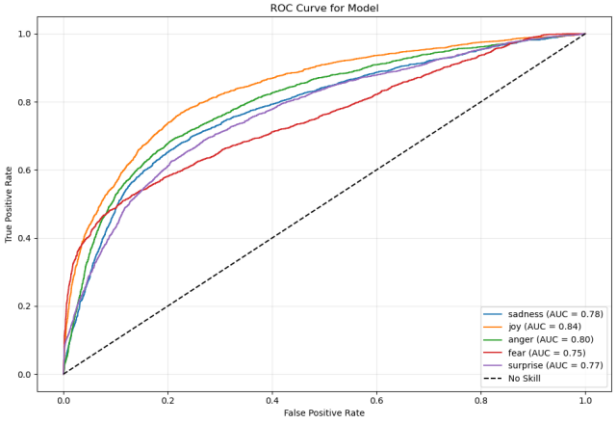
4.7 Emotion classifier Evaluation

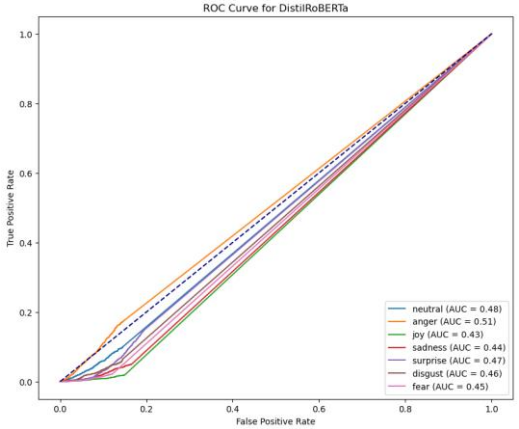


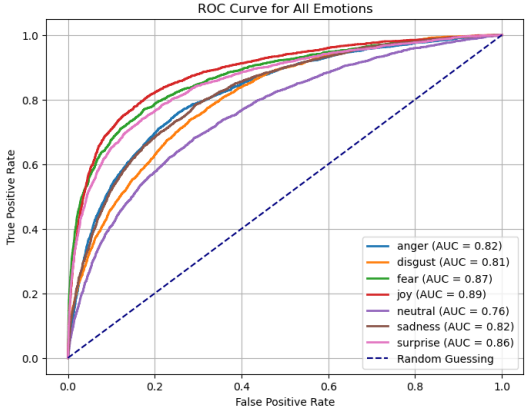
J- Hartman/emotion-English-distilroberta often classify phrases that **reference health concerns or neglect of safety as likely to invoke fear**. In this case, the sarcasm about hygiene and the negative implications for health could trigger an emotional response of fear in sentiment analysis models, even though the broader sentiment is negative

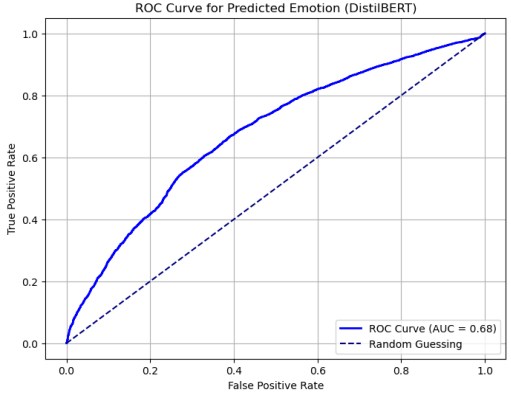
Tweet ID	Sarcasm	Predicted_sentiment	Adjusted_sentiment	Emotion
Swerving two rotten banana skins and a ghastly chewing gum seat, I cleared the coffee cups and moved in to my new home. At Â£40 an hour @TLRailUK was kind enough to bypass hygiene and leave detritus on the floor to play with as itâ€™s always delayed. #CoronaVirusUpdates https://t.co/L8Q1kRt8SJ	1	Negative	Positive	Fear

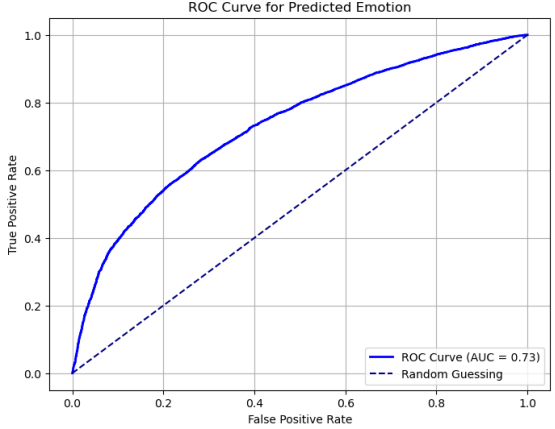
4.8 Emotion classifier Evaluation

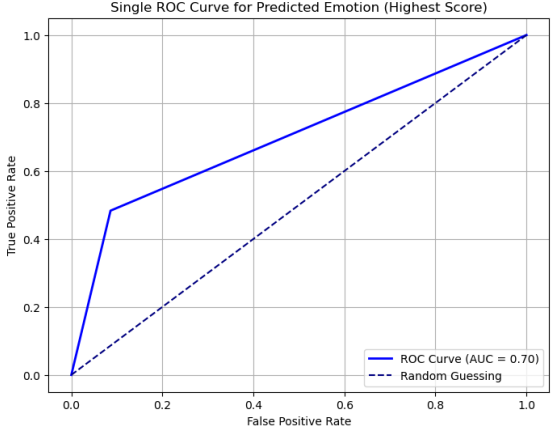












4.9 Emotion classifier Evaluation- Quantitative

Category	DistilRoBERTa	DistilBERT	Michellejieli Emotion Classifier
Model Performance	- Best overall performance among the three, considering **overall ROC AUC (0.73)**	- **Decent performance but struggles with ROC AUC** and precision.	- Performs well for individual emotions but slightly **lower overall score (0.70).**
	- **High Accuracy: 59.25%,** F1-score: 0.59.	- **Moderate Accuracy: 50.23%,** F1-score: 0.47.	- **Low Accuracy: 48.33%,** F1-score: 0.49.
	- High precision: 0.60.	- Moderate precision: 0.56.	- Precision struggles: 0.55.
Confusion Matrix	- Handles emotions like **joy** and **neutral** well, with higher predictions in correct categories.	- Handles **sadness** and **anger** better but misclassifies other emotions.	- Misclassifies most emotions heavily, especially **neutral** and **fear**.
	- **Predicts anger, disgust, and surprise consistently.**	- **Struggles to balance classification across emotions**	- **Strongly biased predictions across all classes.**
ROC Curves	- Lower ROC AUC for individual emotions because DistilRoBERTa only predicts the **highest emotion**, assigning **0 probability to other emotions** by default during ROC curve calculation.	- Moderate ROC AUC, with decent scores for most emotions.	- Strong ROC AUC for individual emotions, indicating balanced performance.
	- **Overall ROC AUC (0.73)** was prioritized due to the inherent limitations of ROC curves for individual emotions.		
Key Takeaways	- Strong overall performer, with balanced predictions and the highest overall ROC AUC.	- Performs moderately well but struggles with classification for less common emotions.	- Good classification for individual emotions, but slightly lower overall performance compared to DistilRoBERTa.

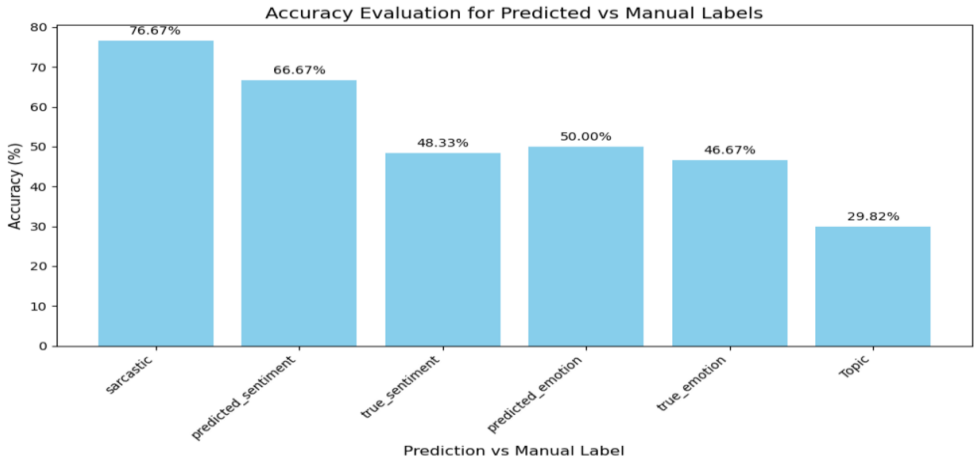
Please find the script here: [Emotion Classifier Evaluation Script]

4.10 Manual Evaluation

We all took 10-10 entris and manually labelled the columns and compared with modeling outputs

[Evaluation CSV]

The accuracy decreases **because errors in sarcasm detection affect sentiment analysis**, which in turn impacts emotion detection. **Each model builds on the previous one, and inaccuracies propagate through the process.**



Model automated annotations

Human Annotations

Scalability

Reduced human labour

More accurate by catching naunaces and understanding context

Subjective criateria

Handles complexity well and understands human tone in a better way

Hochschule für Technik und Wirtschaft Berlin
University of Applied Sciences

4.11 Qualitative Analysis

This experiment involved uploading sarcastic and non-sarcastic tweets to ChatGPT to analyze whether the key performance indicators (KPIs) derived from the content differ based on tone. By comparing the KPIs emerging from sarcastic versus straightforward tweets, we aimed to uncover whether sarcasm introduces unique concerns or merely amplifies existing issues.

Non-Sarcastic Dataset: 13,181 rows

```
kpis = {
    'Punctuality & Reliability': ['delay', 'time', 'late', 'cancel'],
    'Customer Experience': ['service', 'staff', 'experience', 'friendly', 'helpful'],
    'Accessibility': ['barrier', 'access', 'wheelchair', 'disabled'],
    'Facilities & Amenities': ['rubbish', 'wifi', 'toilet', 'clean', 'signage'],
    'Communication': ['process', 'information', 'announcement', 'signs', 'updates'],
    'Complaint Handling': ['repay', 'process', 'issue', 'complaint', 'resolve'],
    'Value for Money': ['expensive', 'cost', 'charge', 'value', 'price'],
}
```

Sarcastic Dataset: 2,568 rows

```
kpis = {
    'Punctuality & Reliability': ['delay', 'time', 'late', 'cancel'],
    'Customer Experience': ['staff', 'service', 'experience', 'helpful'],
    'Accessibility': ['disabled', 'access', 'wheelchair'],
    'Facilities & Amenities': ['toilet', 'wifi', 'rubbish', 'clean'],
    'Communication': ['signs', 'announcement', 'updates', 'information'],
    'Complaint Handling': ['issue', 'repay'],
    'Value for Money': ['cost', 'price'],
}
```

Why It's Interesting:

1. Consistency Across Tones: The experiment showed that ChatGPT identified the same KPIs from both sarcastic and non-sarcastic tweets, suggesting that user tone does not alter the core themes of feedback.

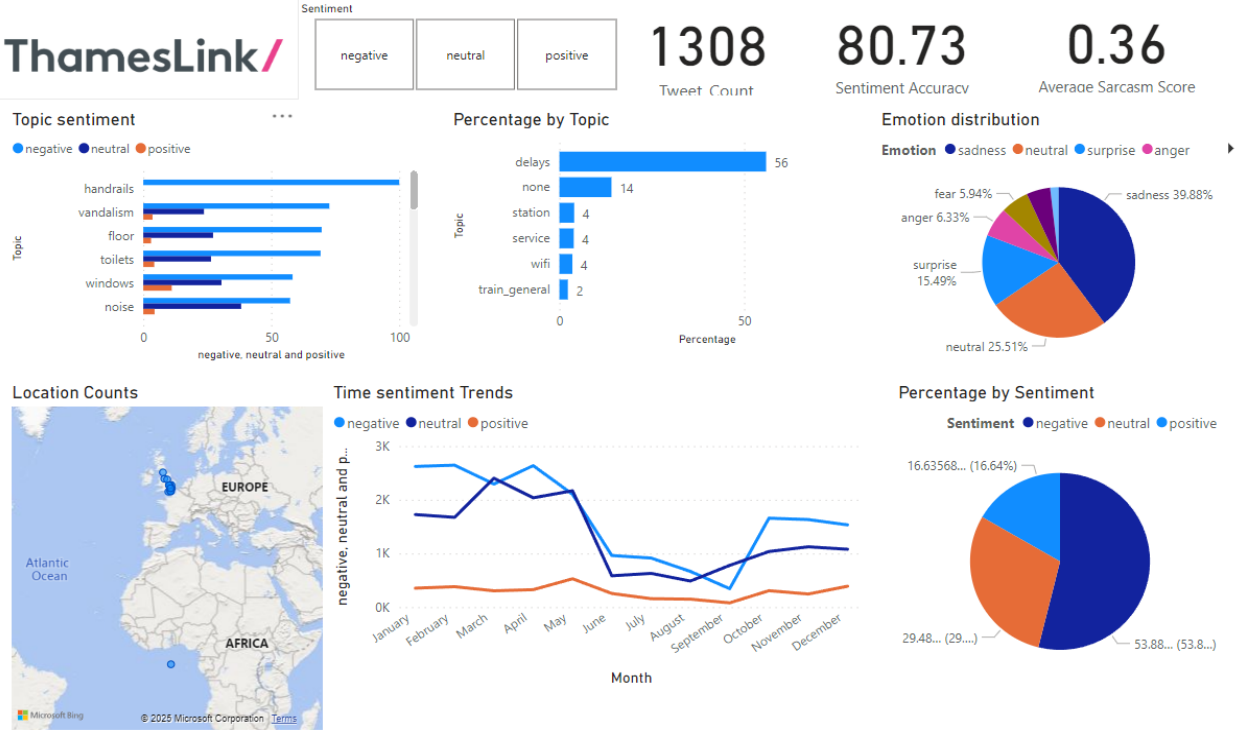
2. Sarcasm as an Emotional Amplifier: While the issues remained the same, sarcasm highlighted areas of heightened dissatisfaction, providing clues about emotionally charged pain points.

3. Actionable Insights with AI: This study demonstrates how AI, like ChatGPT, can consistently extract meaningful insights regardless of tone, helping businesses focus on core themes while recognizing the emotional weight of sarcastic feedback.

Deployment

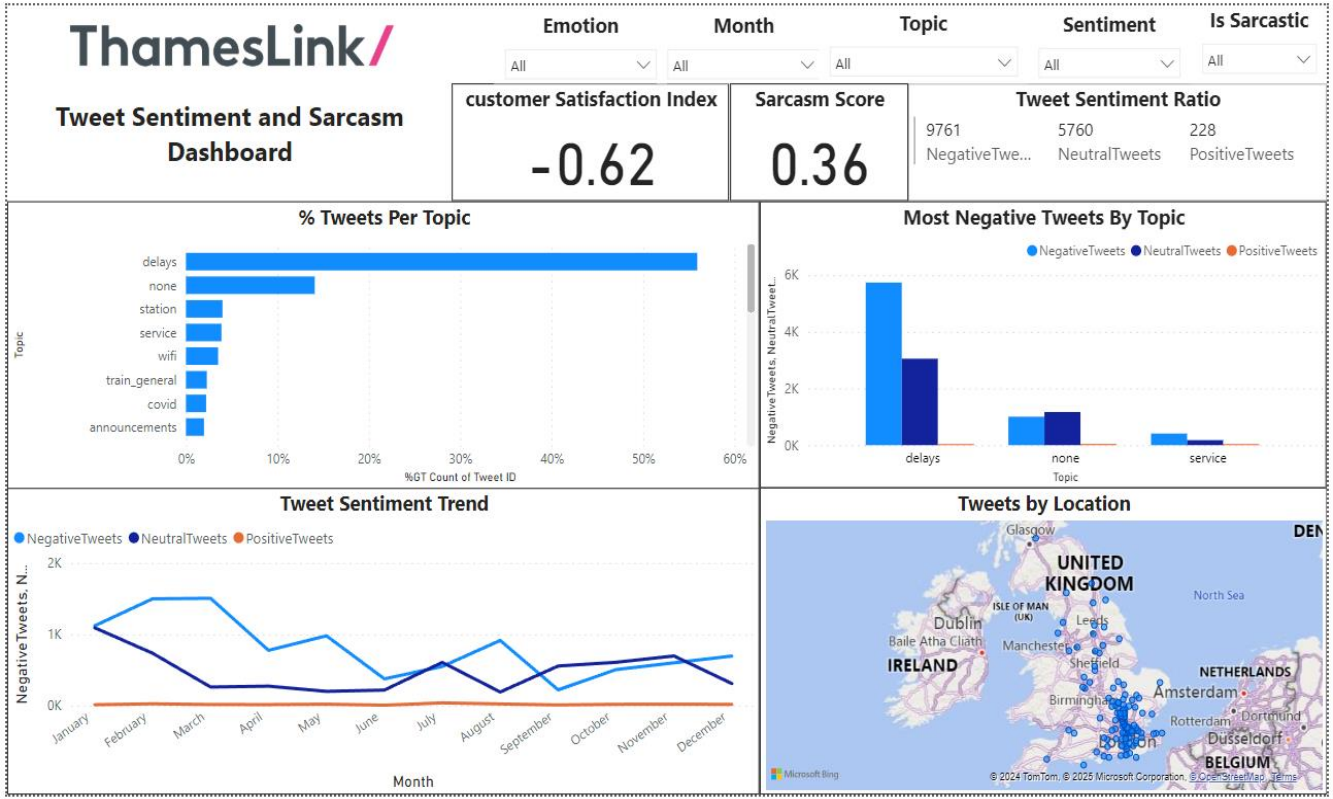
5.1 Our Product – A Tweet Analysis Dashboard Contd..

The Dashboard represents one of the deployment concepts where all the KPIs are represented as a visualization to help Thameslink to reduce the problem of **vehicle availability** by looking at the sarcastic comments posted for it. The KPIs such as **percentage by topic, topic sentiment** shows delays as the main concern among the customer where the KPIs emotion distribution shows the customers feelings through these comments. Visualizations **Percentage by sentiment** shows the distribution of comments in positive, negative and neutral and the **Time sentiment trends** represent the rise and dip of these sentiments. Moreover, **Location count** showcases the location of all the 1308 tweets and the models providing the accuracy of 80.73% with the **average sarcasm score** of 0.36.



5.2 Our Product – A Tweet Analysis Dashboard

The Dashboard represents the overall analysis of the Sarcastic comments where the visualization represents tweet per topic and most negative tweet by represents topic as one of the measure concerns for Thameslink. T**weet sentiment trend** showcases the tweets over the years and **location by tweet** representing the location of these tweets. **Customer satisfaction index** indicates –0.62 which should be **close to 0 to 1**. The filters on the right can be used to filter out the entire dashboard as per any **Emotion** (the interpretation method). The further filters such as **Month, Topic and Sentiment** can be used to filter out the dashboard as per the requirement from the Thameslink. Furthermore, the filter **Is Sarcastic** can be used by Thameslink to filter out only sarcastic comments which can make the decision-making process of the Thameslink team easier.



5.3 KPI Mapping and Suggested Actions

Input

tweet_id	text	sarcastic
1E23FF3	Train P678 was delayed and the toilets were dirty	0

Output

Tweet_id	text	sarcastic	Facilities & Amenities	Punctuality & Reliability	Customer Experience	Accessibility	Communication	Complaint Handling	Value for Money	actions
1E23FF3	Train P678 was delayed and the toilets were dirty.	0	0	1	0	0	0	0	0	Revise and optimize train schedules to improve punctuality
1E23FF3	Train P678 was delayed and the toilets were dirty.	0	1	0	0	0	0	0	0	Upgrade facilities such as toilets, seating, and WiFi access.

5.4 Merits of Using KPI Mapping

Pinpointing Issues

Identifies recurring feedback trends (e.g., "Train delayed") to uncover root causes like scheduling or fleet readiness, aiding optimized resource allocation.

Improved Planning:

Categorizes facility-related feedback (e.g., "Toilets were dirty") into maintenance KPIs, streamlining fleet maintenance for timely availability.

Actionable Insight

Converts unstructured feedback into clear actions (e.g., "Review schedules," "Upgrade WiFi"), accelerating decisions and enhancing efficiency.

Real-Time Monitoring

Connects insights to live dashboards, enabling proactive responses to emerging issues (e.g. "cancellations") to minimize disruptions.

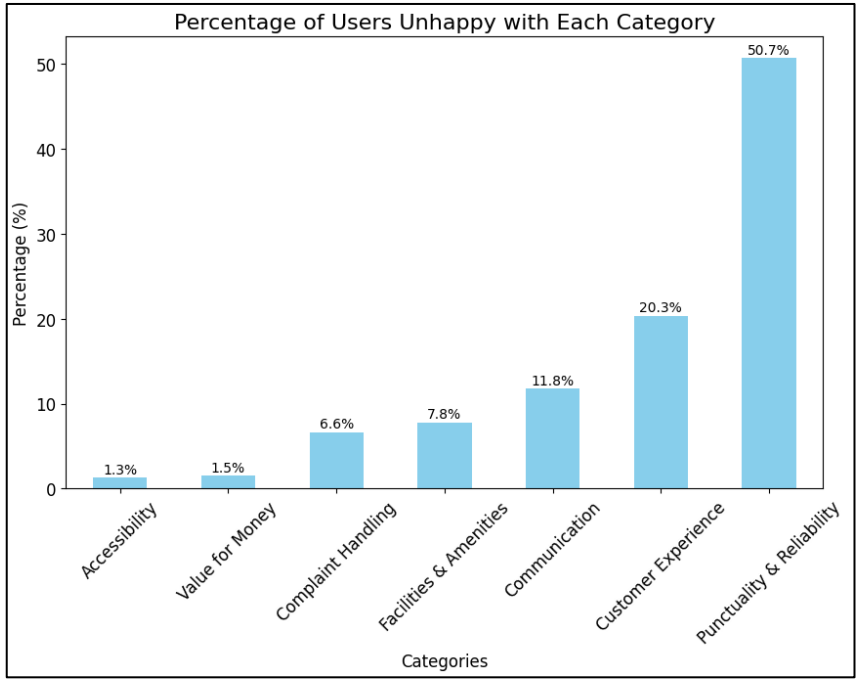
Enhanced Communication

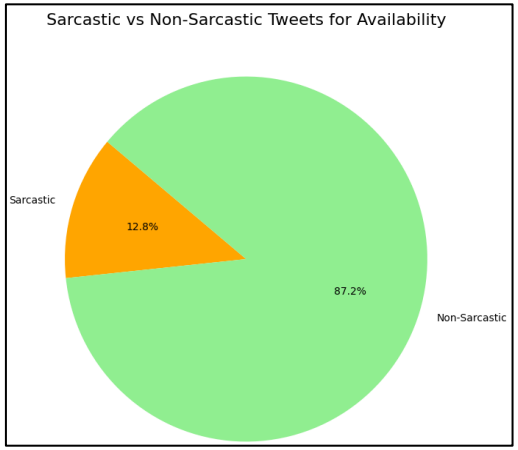
KPIs, prompting solutions like real-time app updates, improving user experience and reducing frustratios.

5.5 Visualizations for KPI Mapping

Tweets are mapped to a Category using a pre-defined set of key words. If user complained about multiple categories, same tweetID is flagged twice, once for each category. This enables us to understand more deeply about the percentage of users dissatisifed with Thameslink's service.

As indicated by the visualizations, 50.7% users were not happy with Vehicle availability, out of these complaints, 12.8% were sarcastic.

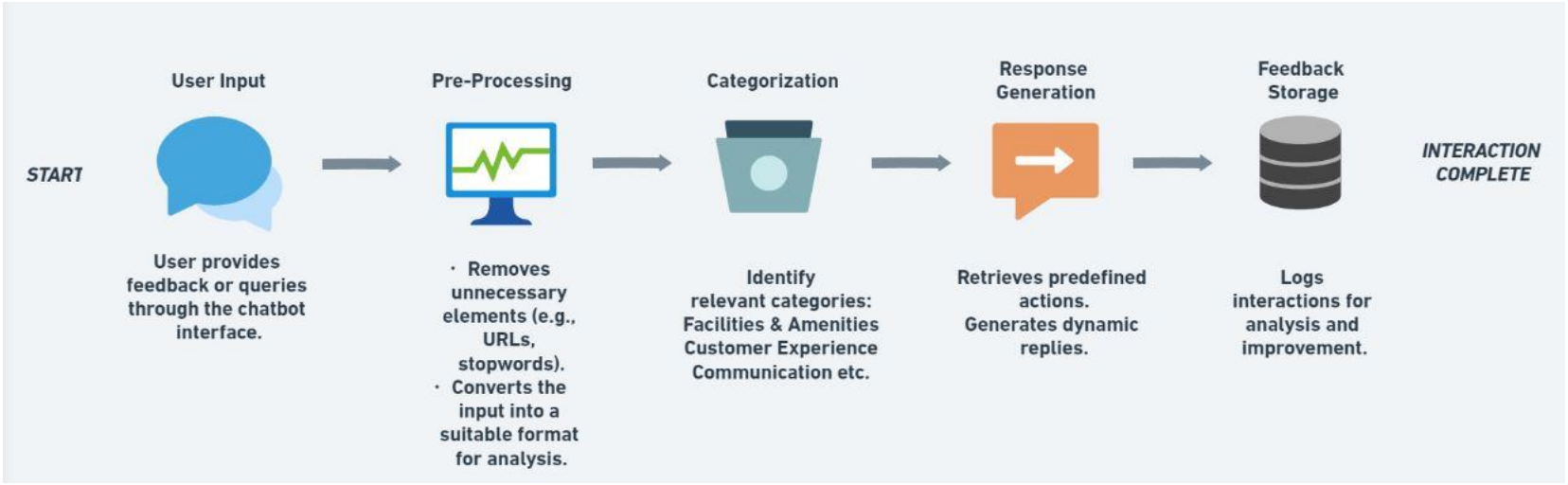




Click Here for Python Notebook

Hochschule für Technik und Wirtschaft Berlin
University of Applied Sciences

5.6 Chatbot Concept



tweet_id	text	s	Facilities &	Punctualit	Customer	Accessibili	Communi	Complaint	Value for I	actions
acd7673f-	the thame	0	1	0	0	0	0	0	0	Upgrade facilities such as toilets, seating, and WiFi access.
5b92aba8	loving the	1	0	0	0	0	0	1	0	Streamline complaint resolution processes and follow up with customers. (Detected as sarcastic: ensure iss

Key Features

- Improves the quality and usability of social media feedback.
- Enables accurate classification and actionable insights.
- Feedback loops ensure the chatbot evolves over time.

Impact

- Enhance vehicle availability by addressing key user concerns.
- Improve customer satisfaction through targeted actions.
- Create a scalable system for handling real-time feedback.

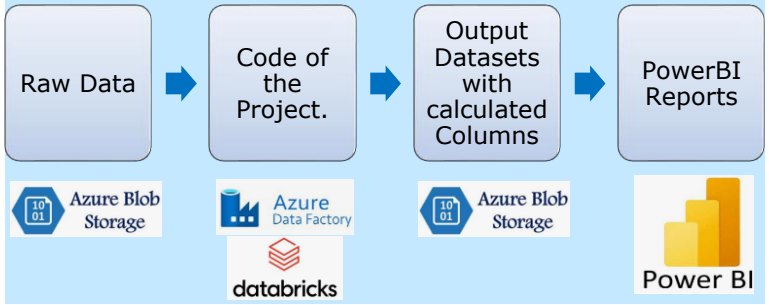
5.7 Recommendations to Thameslink - How to manage data pipeline

Microsoft Azure for Seamless Automation and Scalability

Microsoft Azure

- Automate end-to-end sarcasm detection pipeline with Azure ADF.
- Leverage Azure Blob Storage for centralized and secure data storage.
- Integrate Power BI for real-time visualization and reporting.

WorkFlow



Aspect	Current Approach	Proposed Azure Approach
Workflow Automation	Manual execution in Jupyter Notebook	Fully automated via ADF Pipelines
Data Storage	Local or scattered storage	Centralized in Blob Storage
Scalability	Limited to local resources	Scales with Azure infrastructure
Visualization	Manual export/import to Power BI	Direct integration with Power BI
Real-Time Insights	Challenging	Enabled through event-driven triggers

Why Choose this approach?

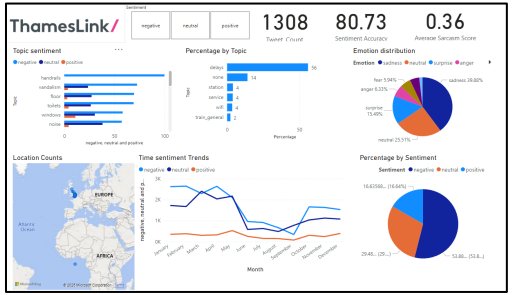
•**Scalability:** Handle growing data volumes with Azure's cloud capabilities.
•**Reliability:** Minimize human error with automated workflows.
•**Actionable Insights:** Power BI dashboards offer quick, interactive insights.
•**Future-Proofing:** Easily integrate new models and features as needs evolve.

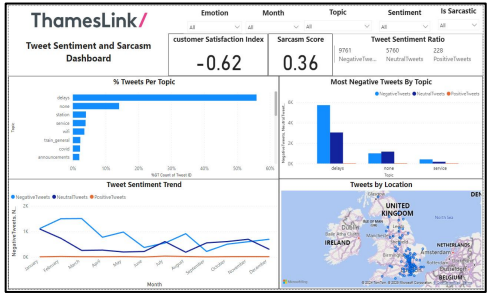
htw Hochschule für Technik und Wirtschaft Berlin
University of Applied Sciences

6.1. Results

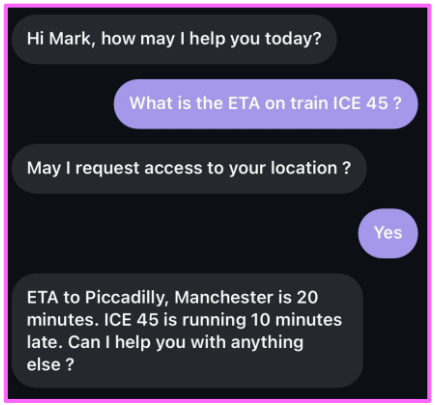
Key Results:

- Sarcasm Detector Model – **RoBERTa Accuracy 82%.** Best performer overall with high accuracy, balanced F1 scores, good recall, and strong ROC AUC.
- Sentiment Detection Model - **Cardiff NLP Twitter RoBERTa . High accuracy (74%)** with the **best ROC AUC across all classes**, offering balanced and robust classification performance.
- Emotion Classifier model – **DistilRoBERTa.** Strong overall performer, with balanced predictions and the highest overall **ROC AUC (0.73), High Accuracy (59.25%).**
- **KPI Mapping with Suggested Actions.**
- Summarization model- we decided to discard this approach due to poor performance.



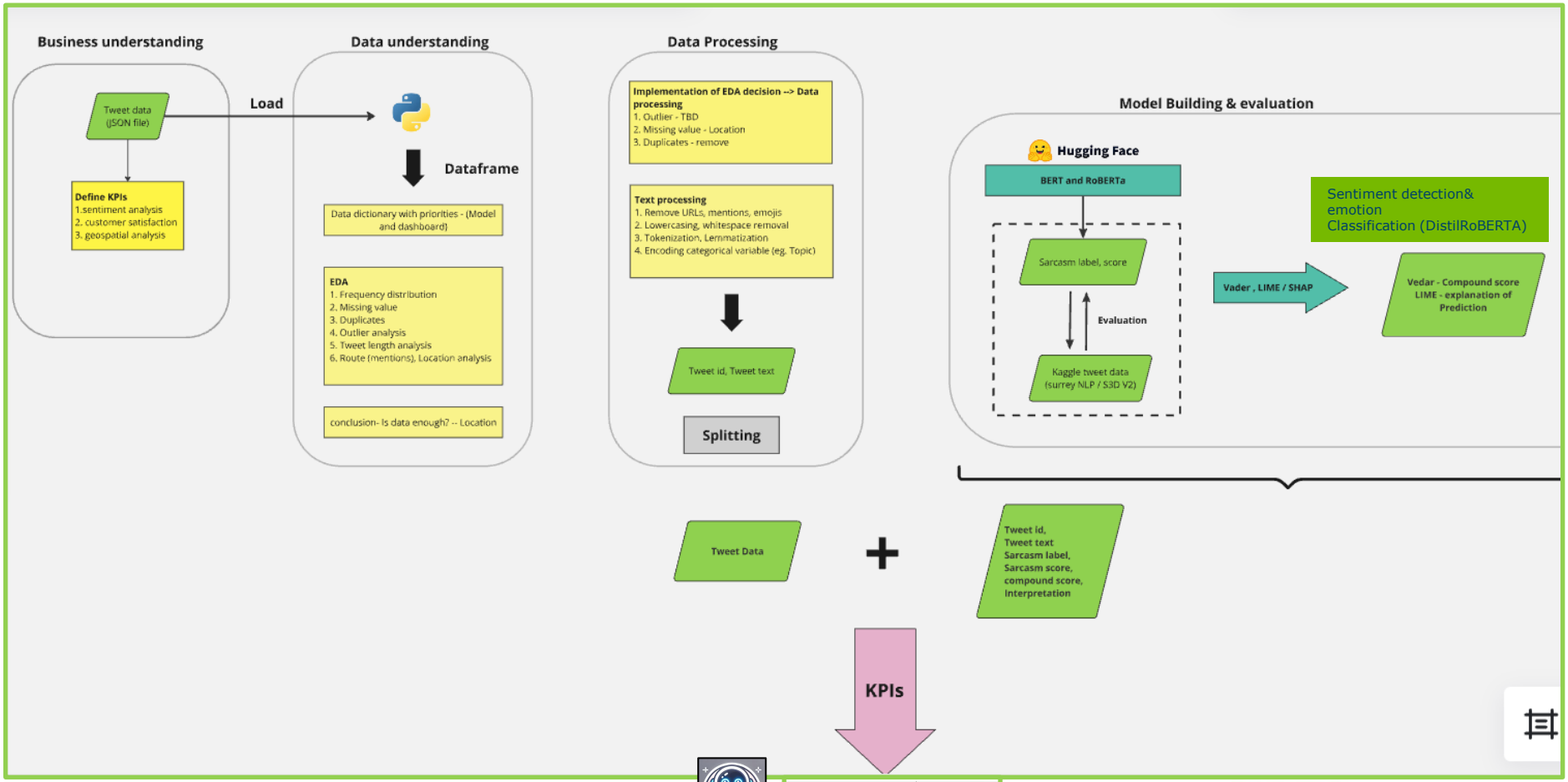


PowerBI Dashboard



ChatBot

7.1. Process



LinkBot

tableau Power BI

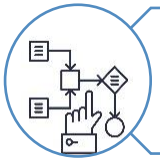
Microsoft Azure

University of Applied Sciences

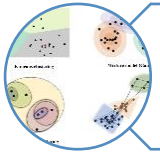
8.1. Executive Summary

Objective: “Analyze tweet data to identify sarcasm, sentiment, and emotions using machine learning and unsupervised learning techniques to improve **vehicle availability**.“

Key Methods Used



Prebuilt algorithms: sarcasm, sentiment and emotion classification



Unsupervised learning: for classifying and analysing tweet data without annotations

Real time visualization: for treads, user and issue prioritization

Deployment concept: user interaction, real time data and issue detection

Key Outcomes

Sarcasm detection: identified sarcasm witth sarcasm scores

Emotion classification: classified emotions like fear, sadness, anger

Sentiment detection: detect sentiment and reverse if sarcasm is present

Actions fro KPI mapping: Qualitative analysis and suggest actions for management to handle relevant issue to KPI

Interactive dashboard: Real time insights in terms of suggested KPIs

Live user tracking: deployment enables real time tracking, with chatbot requesting live location of user for immidiate (daily) insights and improved vehicle availability

8.1. Executive Summary (cont.) : **Recommendations for Thameslink**

- **Azure and Cloud Systems**: Enable scalable modeling pipelines and dashboards to optimize vehicle allocation and ensure real-time visualization of key metrics like availability and user satisfaction.
- **Focus on KPIs**: Monitor vehicle availability, downtime, and user satisfaction in real time for trend analysis, anomaly detection, and proactive resource allocation.
- **Weekly/Monthly Actions**: Regular evaluations of vehicle usage and maintenance schedules optimize fleet performance, reduce bottlenecks, and enhance user experience.
- **User Insights and Dashboard Filters**: Analyze data and feedback to align vehicle availability with passenger demand. Advanced filters offer actionable insights for better prioritization.
- **Chatbots for Monitoring**: Automate vehicle tracking, engage with passengers in real time, and provide updates to boost user satisfaction and operational efficiency.
- **Accurate Data Collection**: Chatbots gather location and feedback, enabling targeted assistance and improved vehicle allocation.
- **Issue Handling**: Align KPIs with actionable tasks to prioritize passenger needs, address concerns, and maintain responsive, reliable service.

9.1. Overall Risks and Mitigations

Risk	Mitigation
1. Importing the data and model output to Power BI every time.	1. Use cloud-based solutions to automate the pipeline and store output as blob storage.
2. Expiry of project management tool Asana.	2. Import the whole project to a new account or renew the tool subscription.
3. Integration and dependency problems in terms of Flask integration and API key configuration while launching chatbot.	3. Host on a cloud-based system instead of a local system.
4. Exposing user data during deployment could compromise privacy or security, especially when handling sensitive data.	4. Use encryption for data at rest and during transit (e.g., AWS KMS or Azure Key Vault).
5. The model misinterprets sarcasm or sentiment, leading to poor user experience or loss of trust.	5. Monitor the model's performance and gather user feedback for continuous improvement. Human in Loop.
6. Domain-specific performance issues due to limited exposure to domain-specific language.	6. Fine-tune models on the dataset to improve domain-specific performance.
7. Emotion Granularity: Emotion classifiers might oversimplify complex emotions, leading to loss of nuanced feedback.	7. Combine multiple models to capture nuanced emotional feedback. Validate results with human annotations.
8. Running large-scale models on big datasets could strain resources, causing delays.	8. Run models on GPU-accelerated systems or scalable cloud platforms. Use efficient models like DistilBERT.
9. Difficulty explaining model classifications, leading to lack of trust in the results.	9. Use visualization tools to explain classification reasoning and build trust with interpretability techniques.
10. Low-quality modeling due to lack of diverse data.	10. Use new datasets for evaluation and preprocessing on the existing dataset.
11. Abbreviation issues often cause inconsistencies during processing.	11. Standardize abbreviations during preprocessing to minimize ambiguity.
12. Identify top 5 words : we tried to map the topics for tweets but only 300 rows were mapped correctly.	12. Mitigation: we decided to exclude this logic from project scope as it was not adding any valuable insights.

9.2. Lessons Learned

Handling unsupervised data	• Leveraged pretrained models to perform sarcasm, sentiment, emotion classification • Faced challenges with less accuracy compared to labelled data
Sarcasm and sentiment integration	• Reversing sentiments for sarcastic texts improves downstream emotion classification • Clear workflow and interdependencies are essential for maintaining result integrity.
Challenges with non standard language	• Pretrained models for interpretation and summarisation struggled with slang, informal grammer • Domain specific data augmentation and fine tuning are necessary for improved performance
Emotion classification	• Allows business to provide immediate attention to emotionally charged situations • Identify customers experiencing negative emotions eg, frustration, fear to prioritize their issues
Value of iterative management	• Structured project management improves workflow efficiency and improve task division • Helps to focus on area of expertise of group member by assigning role based task
Human-in-the-loop for enhanced accuracy	• Incorporating human annotaions for cross checking the pretrained model generated outputs ansures the reliability and accuracy