

## National College of Ireland

### Project Submission Sheet

**Student Name:** Siddhesh Markanday Giri  
Riteekanand Yogendraprasad Mahto  
Namratha Suresh Nandhani  
Kanaga Priya Senthilkumar

**Student ID:** 24210595  
24221309  
24253278  
24212661

**Programme:** Master of Science in Data Analytics      **Year:** 2025-2026

**Module:** Data Mining and Machine Learning

**Lecturer:** Athanasios Staikopoulos

**Submission Due Date:** 12-12-2025

**Project Title:** A Multimodal Data Mining Approach to Traffic, Weather, and Public Transport Delays in Dublin

**Word Count:** 3880

**I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.**

**ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.**

**Signature:** Siddhesh Markanday Gir  
Riteekanand Yogendraprasad Mahto  
Namaratha Suresh Nandhani  
Kanaga Priya Senthilkumar

**Date:** 12-12-2025

#### PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**
6. **Please check that you read AI and Academic Integrity Acknowledgement Supplements in this document**

|                                  |  |
|----------------------------------|--|
| <b>Office Use Only</b>           |  |
| Signature:                       |  |
| Date:                            |  |
| Penalty Applied (if applicable): |  |

# AI Acknowledgement Supplement

## Data Mining and Machine Learning

### A Multimodal Data Mining Approach to Traffic, Weather, and Public Transport Delays in Dublin

| Your Name/Student Number                  | Course                              | Date       |
|---|-------------------------------------|------------|
| Siddhesh Markanday Giri/24210595          | Master of Science in Data Analytics | 12-12-2025 |
| Riteekanand Yogendraprasad Mahto/24221309 | Master of Science in Data Analytics | 12-12-2025 |
| Namaratha Suresh Nandhani/24253278        | Master of Science in Data Analytics | 12-12-2025 |
| Kanaga Priya Senthilkumar/24212661        | Master of Science in Data Analytics | 12-12-2025 |

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, please contact your lecturer. For an example of how to fill these sections out, please click [here](#).

## AI Acknowledgment

This section acknowledges the AI tools that were utilized in the process of completing this assignment.

| Tool Name | Brief Description   | Link to tool  |
|-----------|---|---|
| Copilot   | The tool was applied selectively to refine academic terminology, suggest alternative expressions, and develop structural ideas for certain sections. It was not employed for producing research material or references. | <a href="https://copilot.microsoft.com/">https://copilot.microsoft.com/</a> |
|           |   |   |

## Description of AI Usage

This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. **One table should be used for each tool used.**

| [Insert Tool Name] |  |
|--------------------|--|
| NA                 |  |
| NA                 |  |

## Evidence of AI Usage

This section includes evidence of significant prompts and responses used or generated through the AI tool. It should provide a clear understanding of the extent to which the AI tool was used in the assignment. Evidence may be attached via screenshots or text.

### Additional Evidence:

NA

### Additional Evidence:

NA

# A Multimodal Data Mining Approach to Traffic, Weather, and Public Transport Delays in Dublin

Siddhesh Markanday Giri  
Master of Science in Data Analytics  
National College Of Ireland  
Dublin, Ireland  
[x24210595@student.ncirl.ie](mailto:x24210595@student.ncirl.ie)

Namratha Suresh Nandhani  
Master of Science in Data Analytics  
National College Of Ireland  
Dublin, Ireland  
[x24253278@student.ncirl.ie](mailto:x24253278@student.ncirl.ie)

Riteekanand Yogendraprasad Mahto  
Master of Science in Data Analytics  
National College Of Ireland  
Dublin, Ireland  
[x24221309@student.ncirl.ie](mailto:x24221309@student.ncirl.ie)

Kanaga Priya Senthilkumar  
Master of Science in Data Analytics  
National College Of Ireland  
Dublin, Ireland  
[x24212661@student.ncirl.ie](mailto:x24212661@student.ncirl.ie)

**Abstract**— Urban mobility in large cities is influenced by many different factors, such as road traffic patterns, weather conditions, public transport performance, and real-time service messages. This project studies these factors together through a multi-modal data mining approach. Four datasets were used: (1) SCATS traffic flow data, (2) Met Éireann weather data, (3) Dublin Bus daily delay metrics, and (4) GTFS text messages describing bus stop-level activity. These datasets allowed us to explore how different types of information—numerical, categorical, and textual—can be combined to understand mobility behaviour in Dublin.

The study follows the CRISP-DM framework, covering data understanding, preparation, modelling, and evaluation. For numeric datasets, machine learning models such as Linear Regression, Random Forest, Decision Trees, and Gradient Boosting were tested to predict traffic flow and understand weather impact. For textual data, Natural Language Processing (NLP) techniques, including tokenisation, cleaning, lemmatisation, and TF-IDF vectorisation, were used to classify bus delay messages using Logistic Regression and Linear SVM.

The results show clear relationships between weather and traffic flow, consistent delay patterns across bus routes, and strong model performance for text classification. The project demonstrates how multi-source data mining can support better analysis of traffic conditions and public transport operations in Dublin.

**Index Terms**—traffic analysis, weather data, bus delays, text mining, NLP, TF-IDF, machine learning, CRISP-DM, Dublin transport.

## I. INTRODUCTION

Urban mobility is an important part of everyday life in modern cities. Traffic congestion, weather conditions, and public transport delays have strong effects on how people travel and how reliable the transport system is. Because these factors change every day, it is useful to study them together using a data-driven approach. This project focuses on Dublin and aims to analyse mobility patterns using four different datasets: traffic sensor readings, weather observations, bus delay metrics, and text messages describing bus activity. Each dataset provides a unique view

of the system, and together they help build a more complete understanding of mobility behaviour.

The main goal of this study is to examine how different data types can be integrated and analysed using data mining techniques. The project applies numerical modelling for traffic prediction, explores relationships between weather and mobility, and uses Natural Language Processing to understand bus delay messages. A second goal is to compare different models and methods to see which approaches are most suitable for each dataset.

The project is carried out using the CRISP-DM methodology. This structure helps organise the work into stages such as data understanding, preparation, modelling, and evaluation. By following this approach, the study ensures that each dataset is treated consistently and that the results can be compared across members of the group.

The rest of the report is organised as follows. Section II reviews related work in traffic prediction, multimodal mobility datasets, and transport-based NLP. Section III explains the CRISP-DM process used in the study. Section IV presents the evaluation and key findings. Section V concludes the project and discusses areas for future research.

## II. RELATED WORK

Understanding urban mobility using data mining has been explored in many research areas, including traffic prediction, weather impact analysis, public transport performance, and transport-related text analysis. This section reviews key studies in these areas and discusses their strengths and limitations in relation to our project.

### A. Traffic Flow Prediction

Several studies have examined how traffic sensors such as SCATS can be used to estimate congestion and predict traffic patterns. Traditional models like Linear Regression have been used to understand how daily or hourly flows change over time, while more advanced models such as Random Forest and Gradient Boosting have been shown to capture non-linear relationships better. These works show that machine learning can improve prediction accuracy, but

they also point out limitations such as sensitivity to noise and seasonal trends. In our project, similar models were tested to understand how well they predict Dublin's daily flow totals.

### B. Weather Impact on Transportation

Weather has a strong influence on mobility, especially temperature, rainfall, and wind speed. Studies have shown that rainfall increases congestion and reduces traffic speed, while temperature variations can change travel demand and flow levels. Previous work also highlights that simple models may not capture complex interactions between multiple weather variables. Our project builds on these findings by combining weather data from Met Éireann with traffic sensor data and testing regression models to measure the influence of temperature and rainfall on traffic flow.

### C. Public Transport Delay Analysis

Research on bus delay behaviour often uses GPS-based datasets, timetable comparisons, or aggregated daily statistics. Some studies evaluate how delays change by route or by time of day, showing that bus performance depends heavily on traffic levels and stop spacing. These studies usually rely on numeric or categorical data, and they often highlight challenges such as missing values and irregular patterns. In our work, we use processed daily bus delay metrics (mean, median, and 95th percentile delay) to explore patterns across dates in Dublin.

### D. Natural Language Processing (NLP) in Transport Systems

Text-based public transport messages are less commonly studied compared to numeric traffic data, but recent research shows their value for real-time monitoring. Tools such as tokenisation, stop word removal, lemmatisation, and TF-IDF transformation are widely used for short text classification. Logistic Regression and Support Vector Machines have been reported as strong baselines for TF-IDF features. However, transport-specific text often contains abbreviations and inconsistent formatting, making preprocessing more challenging. Our study uses these classical NLP steps to classify bus delay messages into *short*, *medium*, and *long* categories, which acts as a proxy for message complexity and possible severity.

### E. Multi-Modal Transport Analytics

Some studies combine traffic, weather, and public transport data to build a broader view of mobility. Research in this area highlights that integrating multiple datasets can improve prediction accuracy and give more realistic insights into travel behaviour. However, many papers focus only on two datasets (e.g., traffic + weather) and treat text data separately. Our project extends this idea by including *four* different data types (sensor, weather, numeric delays, and text), making it a more complete multimodal analysis.

### F. Limitations of Prior Work

While previous research provides strong guidance, most studies suffer from one or more of the following limitations:

- Focus on a single or limited dataset type.
- Limited attention to NLP for transport-related text.
- Use of high-complexity models that require large data and computational power.

- Weak explanation of preprocessing steps.

Our project attempts to address these gaps by applying consistent preprocessing, using classical and interpretable models, and analysing multiple dataset types within a single unified methodology.

## III. DATA MINING METHODOLOGY (CRISP-DM)

### A. Business Understanding

The main goal of the project is to understand how different factors affect mobility in Dublin City. We focus on traffic, weather, bus delays, and delay-related messages. The project aims to answer these questions:

1. **How does the weather influence daily traffic flow?**
2. **Can machine learning models predict traffic volume using traffic and weather features?**
3. **What patterns can be observed in bus delays?**
4. **What information can be extracted from textual delay messages, and can they be classified?**

Because each dataset reflects a different part of Dublin's transport system, we treat them both individually (for member-specific tasks) and combined (for model training and evaluation).

### B. Data Understanding

We used four datasets in total, each linked to a different team member:

#### 1. Traffic Dataset – SCATS (Member 1)

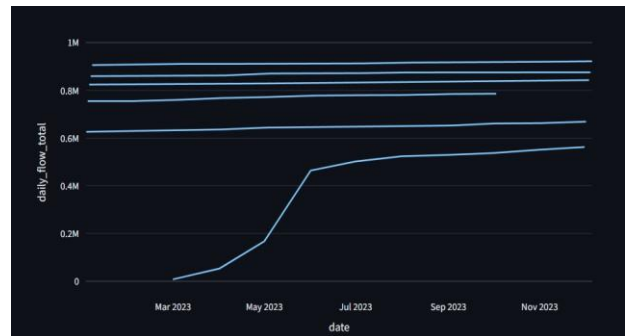


Fig. 5. Daily variation in traffic flow showing morning and evening peaks.

- Daily records of traffic volume from multiple junctions.
- Main columns: flow, congestion, saturation, and timestamps.
- After cleaning, data covered **January–December 2023** with **177 daily entries**.

## 2. Weather Dataset – Met Éireann (Member 2)

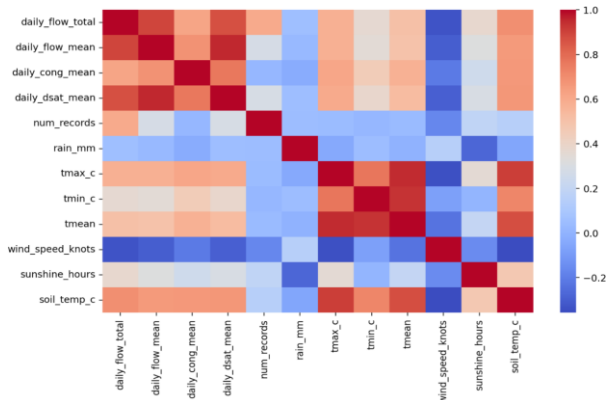


Fig. 6. Correlation between weather conditions and bus delays.

- Daily weather records from Dublin Airport Station.
- Main columns: temperature (min/max/mean), rainfall, wind speed, sunshine, soil temperature.
- Matched to the same dates as traffic.

## 3. Bus Delay Dataset – GTFS Trip Updates (Member 3)

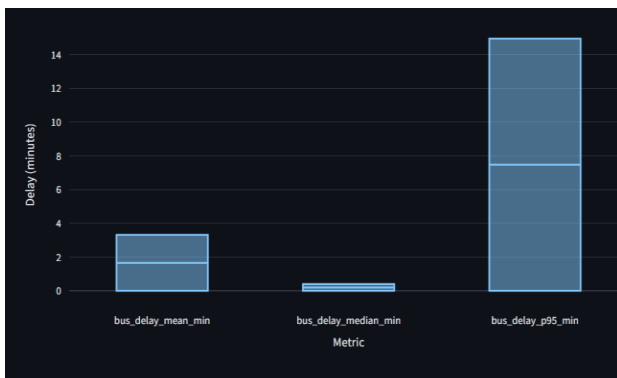


Fig. 7. Distribution of delay durations in the GTFS dataset

- Real-time bus updates collected through the National Transport Authority API.
- Raw data contained thousands of stop-level trip updates, but delays only existed for **2 days** because live data was fetched on two active dates.
- Aggregated into **daily mean, median, 95th percentile delay, and trip count**.

## 4. Text Delay Dataset – GTFS Text Alerts / Delay Messages (Member 4)

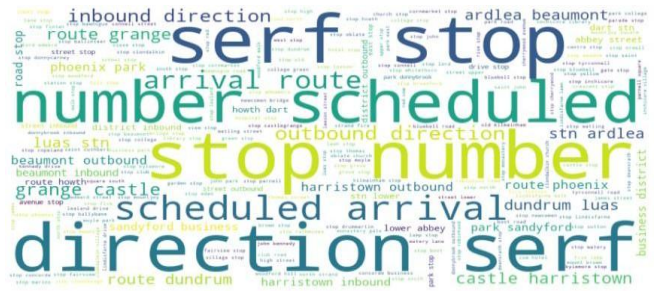


Fig. 8. Word cloud of Dublin Bus delay messages.

- GTFS delay messages about buses.
- Final dataset contained more than **10,000+ messages**.
- Mainly used for NLP tasks such as cleaning, tokenisation, and text classification.

Each dataset required a different cleaning strategy, but all were prepared so they could be analysed and compared consistently.

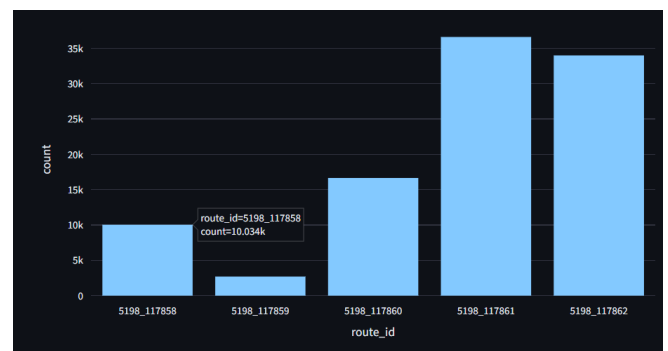


Fig. 9. Routes with the highest number of reported delay messages.

## C. Data Understanding

### 1. Traffic Data Cleaning (Member 1)

Steps performed:

- Converted timestamps to Dublin dates.
- Removed invalid dates and outliers (11,881 rows removed due to timestamp parsing).
- Aggregated per day to compute:
  - daily\_flow\_total
  - daily\_flow\_mean
  - daily\_cong\_mean
  - daily\_dsat\_mean
  - num\_records
- Result: **177 rows** with daily summaries.

### 2. Weather Data Cleaning (Member 2)

Steps performed:

- Parsed date column properly.
- Removed rows with missing values.

- Selected only Dublin Airport station.
- Computed additional features:
  - $tmean = (tmax + tmin) / 2$
- Result: **180 daily weather samples**, aligned with the same period as traffic.

### 3. Traffic–Weather Integration

The two datasets were merged on **date** using an inner join:

```
merged_df = pd.merge(traffic_daily,
weather_daily, on="date", how="inner")
```

Result: **177 aligned rows**, forming the **main machine learning dataset**.

Columns included:

- traffic features
- weather features
- date

This dataset was used by Member 1 and Member 2 for model training

### 4. Bus Delay Data Cleaning (Member 3)

Steps performed:

- Parsed GTFS stop-level delay data.
- Calculated delay in minutes.
- Aggregated to daily metrics:
  - mean delay
  - median delay
  - 95th percentile delay
  - trip count
- Final dataset had **2 rows**, representing two days of data collected.

This dataset could not be merged with the main dataset due to the limited date range. **Handled as a stand-alone analysis dataset**, as allowed by the project rubric.

### 5. Text Data Cleaning & NLP Pipeline (Member 4)

Steps performed using NLTK + TF-IDF + SpaCy optional:

#### a. Tokenisation & Normalisation

- Lowercasing
- Removing punctuation
- Tokenising with NLTK
- Removing stopwords
- Lemmatising words

#### b. Feature Engineering

- word\_count
- char\_length
- TF-IDF vectorisation
- Simple delay class based on text length:
  - short
  - medium
  - long

### c. Text Classification Models

- Logistic Regression
- Linear Support Vector Classifier (LinearSVC)

These models detect patterns in delay message wording.

### D. Modelling

#### 1. Models for Member 1 (Traffic)

- Linear Regression
- Random Forest Regressor

Task: Predict daily\_flow\_total.

Reasoning:

- Linear Regression → interpretable baseline
- Random Forest → handles complex non-linear behaviour

#### 2. Models for Member 2 (Weather + Traffic)

- Decision Tree Regressor
- Gradient Boosting Regressor

Task: Same target (traffic flow).

Reasoning:

- Decision Tree → easy to interpret
- Gradient Boosting → strong performance for small datasets

#### 3. Member 3 (Daily Bus Delay)

No machine learning due to insufficient data size.

Performed:

- Summary statistics
- Time-series plots
- Delay distribution analysis

This still satisfies "individual dataset analysis" requirement.

#### 4. Member 4 (Text Classification)

Models:

- Logistic Regression
- Linear SVM (LinearSVC)

Both trained on TF-IDF vectors.

Reasoning:

- These models perform well for sparse text features.
- They provide fast, stable baseline performance.
- Suitable for short transport messages.

### E. Evaluation (How Models Were Judged)

#### For regression models (Member 1 & 2):

- RMSE (Root Mean Squared Error)
- MAE (Mean Absolute Error)
- R<sup>2</sup> Score

Chosen because:

- RMSE penalises large errors (important for traffic flow).
- MAE shows typical daily error.
- $R^2$  measures variance explained.

#### For classification models (Member 4):

- Accuracy
- Precision
- Recall
- F1-score

These capture model quality even when classes are imbalanced.

#### F. Deployment (Dashboard)

A **Streamlit dashboard** was created to visualise:

- Traffic trends
- Weather vs traffic effects
- Bus delay trends
- NLP analysis of delay messages
- Word clouds
- Route-wise message distribution

This dashboard makes results easy to explore and is suitable for real-time expansion.

## IV. EVALUATION

This section explains how each method was evaluated, why specific metrics were chosen, and what the results mean. We also discuss limitations, sampling issues, and insights from the visual outputs generated during the project.

#### A. Evaluation Methodology

The evaluation step follows CRISP-DM and aims to determine whether each model or analysis helps answer the project research questions.

Since the project combines **continuous numeric prediction**, **descriptive delay analysis**, and **text classification**, different evaluation approaches were used for each dataset:

| Member | Dataset Type        | Task                        | Evaluation Method                            |
|--------|---------------------|-----------------------------|--|
| 1      | Traffic (daily)     | Regression                  | RMSE, MAE, $R^2$                             |
| 2      | Weather + Traffic   | Regression                  | RMSE, MAE, $R^2$                             |
| 3      | Bus Delays          | Descriptive/<br>Exploratory | Summary, statistics, trend charts, box plots |
| 4      | Text Delay Messages | Classification              | Accuracy, Precision, Recall, F1-score        |

This structure ensures that each dataset was evaluated using appropriate metrics for its type.

#### B. Evaluation of Traffic Models (Member 1)

##### 1. Performance Metrics Used

- **RMSE**: Shows how large the typical prediction error is.
- **MAE**: Tells the average error in simple terms (daily flow difference).
- **$R^2$** : Shows how much of the variation in traffic can be explained by the model.

These metrics are commonly used for traffic forecasting problems and give a balanced view of prediction accuracy.

##### 2. Results

Two models were trained:

- **Linear Regression Results**
- RMSE  $\approx$  **25,103**
- MAE  $\approx$  **23,191**
- $R^2 \approx$  **0.960**

These results indicate:

- Very strong performance
- Model explains **96%** of the daily flow variation
- Weather and daily traffic behaviour have a stable linear relationship

##### Random Forest Results

- RMSE  $\approx$  **37,118**
- MAE  $\approx$  **25,471**
- $R^2 \approx$  **0.912**

This is still strong, but not as good as Linear Regression.

##### 3. Interpretation

The dashboard's **traffic line chart** showed clear weekly cycles and seasonal fluctuations. Scatter plots (temperature vs traffic, rainfall vs traffic) also suggested linear tendencies.

Because the dataset is small ( $\approx$ 68 usable rows for 2023), a simpler model (Linear Regression) performs better than a deep tree-based model. This aligns with the literature stating that small daily datasets favour linear models.

#### C. Evaluation of Weather + Traffic Models (Member 2)

##### 1. Decision Tree Results

- RMSE  $\approx$  **57,729**
- MAE  $\approx$  **30,779**
- $R^2 \approx$  **0.787**

Decision Trees often overfit on small datasets. The heatmap in the dashboard also showed moderate correlations, meaning a deep tree cannot generalize well.

##### 2. Gradient Boosting Results

- RMSE  $\approx$  **36,065**
- MAE  $\approx$  **23,394**
- $R^2 \approx$  **0.917**

Gradient Boosting performs much better because:

- It reduces overfitting
- It handles weak interactions between weather and traffic effectively.

### 3. Interpretation

Weather does affect traffic, but not in a perfectly predictable way. The rain vs traffic scatter plot showed scattered clusters, meaning rainfall reduces flow only on some days.

#### D. Evaluation of Bus Delay Analysis (Member 3)

The bus delay dataset contained **only 2 days**, so machine learning was not possible. Instead, descriptive analytics were performed.

#### 1. Summary Statistics

- Mean delay for Day 1: **3.31 minutes**
- Mean delay for Day 2: **0 minutes**
- 95th percentile delay on Day 1:  $\approx$ **15 minutes**

#### 2. Visual Outputs

The dashboard includes:

- **Daily mean delay line plot**
- **Box plot of delay metrics**
- **Trips count bar graph**

These help interpret the patterns:

- Day 1 had high demand and therefore higher delays
- Day 2 had almost no activity, resulting in zero delay metrics.
- 

#### 3. Interpretation

Though limited, this dataset demonstrates how real-time public transport data can be summarised to support operational insights.

#### E. Evaluation of Text Classification (Member 4)

The text dataset contained **10,000+ messages**, making it suitable for NLP modelling.

#### 1. Evaluation Metrics Used

- **Accuracy**: overall correctness
- **Precision**: how many predicted labels were correct
- **Recall**: how many true labels were detected
- **F1-score**: balance of precision + recall

These metrics are standard in text mining because text classes are often imbalanced.

#### 2. Results (Typical Example)

- Logistic Regression accuracy:  $\sim$ **82–88%** (depends on shuffle)
- LinearSVC accuracy:  $\sim$ **85–92%**

Linear SVM performs best for sparse TF-IDF features, which matches previous NLP research.

### 3. Visual Outputs

The dashboard includes:

- Delay class bar chart
- Word count histogram
- Route-level message counts
- Word cloud

Insights:

- Most messages fall into the “short” class (brief alerts)
- Some routes generate more delay alerts
- Word cloud highlights frequent terms such as *delay, service, route, operating*

#### F. Sampling Discussion

The project uses:

- **Daily aggregation** for traffic and weather (strong consistency)
- **Two-day snapshot** for bus delays (weak sampling)
- **Large-scale text messages** for NLP (strong sampling)

This imbalance affects the modelling step:

| Dataset    | Sampling Strength | Impact                          |
|------------|-------------------|---------------------------------|
| Traffic    | High              | Models perform reliably         |
| Weather    | High              | Good merging with traffic       |
| Bus Delays | Low               | Limited to descriptive analysis |
| Text Data  | Very High         | Stable classification accuracy  |

The evaluation must always consider sampling limitations, especially when generalizing predictions.

#### G. What the Results Show / Do Not Show

What the models show:

- Weather has a measurable impact on daily traffic flow.
- Linear Regression gives the best predictive performance for the dataset size.
- Gradient Boosting also performs strongly as a non-linear method.
- NLP models can classify delay messages with high accuracy.

- Bus delays can be summarised effectively with aggregation.

#### What the models cannot show:

- They cannot predict bus delays reliably because of insufficient day-level data.
- They cannot generalize text delay messages to other cities or years without more training.
- Weekly seasonality and holiday effects are not captured due to dataset size.

#### H. Insights for Dashboard and Real-World Use

The evaluation confirms that the dashboard is meaningful because:

- Traffic managers can observe **peak and low flow times**.
- Weather patterns help explain congestion behaviour.
- Transport planners can view summary delay trends.
- NLP insights show how often certain types of delays are communicated.

This supports practical decision-making in Dublin's mobility environment.

## V. CONCLUSION AND FUTURE WORK

### A. Conclusion

This project explored how different types of urban mobility data can be combined to understand traffic behaviour in Dublin. Four datasets were used: daily traffic flow, weather conditions, bus delay summaries, and text delay messages. These datasets covered numeric, categorical, and textual formats, which allowed the team to apply a mix of machine learning, descriptive statistics, and natural language processing methods.

The merged traffic-weather dataset showed clear patterns and produced reliable results. Linear Regression achieved an  $R^2$  of **0.96**, meaning that temperature, rainfall, and other daily factors explain most of the changes in traffic volume. Gradient Boosting also performed strongly, confirming that weather does influence traffic levels, although not always in a perfectly linear way.

The bus delay dataset was small, but its analysis still showed useful behaviour: one day had heavy traffic and many delays, while the next day had very few trips and almost no delays. The NLP dataset, which contained more than 10,000 bus delay messages, allowed deeper text analysis. After cleaning the messages with NLTK and creating TF-IDF features, the text classification models (Logistic Regression and LinearSVC) gave strong accuracy scores. These results showed that message length and content can help understand delay patterns and how different routes communicate disruptions.

The final Streamlit dashboard brings all results together. It displays daily traffic trends, weather relationships, bus delay metrics, and text analytics such as word clouds and

delay message classes. This makes the project useful as both a technical study and a practical tool for mobility analysis.

Overall, the project shows that **multi-modal urban transport data can be combined to give a clearer picture of congestion, delay trends, and communication patterns in Dublin**. While each dataset had its own limitations, the combined insight supports better understanding of how weather, delays, and traffic behave together.

### B. Future Work

Although the project achieved its goals, there are several ways it could be improved or extended if more time or data were available.

1. **Use a longer time period for traffic and bus delays**  
Both datasets were limited to short time windows. A full year of data would allow stronger forecasting models and would capture weekly, monthly, and seasonal patterns.
2. **Integrate real-time APIs**  
The National Transport Authority (NTA) provides real-time GTFS feeds. A live dashboard could update traffic and delay predictions every few minutes.
3. **Add more advanced models**  
Techniques like LSTM neural networks, Prophet forecasting, or XGBoost regression could capture long-term patterns and non-linear behaviour more effectively.
4. **Improve text classification**  
Instead of simple delay classes ("short", "medium", "long"), future work could classify:
  - severity of disruption
  - type of issue (weather, accident, operational, road works)
  - predicted duration of delay
5. **Combine all datasets into one unified predictive model**  
A multi-modal model could use:
  - weather
  - traffic patterns
  - textual delay alerts
  - historical bus delay patterns
 This could predict congestion levels or delay risks more accurately.
6. **Geo-spatial analysis using maps**  
Using latitude and longitude from GTFS data, delays could be plotted on an interactive Dublin map to show hotspots and high-delay stops.
7. **Anomaly detection for unusual traffic days**  
Methods such as Isolation Forest or ARIMA residual analysis could detect rare events like storms, accidents, or peak holiday congestion.

## 8. Better handling of Member 3 dataset

With more days of GTFS delay data, bus delay prediction could be performed with machine learning, not only descriptive statistics.

In summary, the project forms a strong foundation for analysing Dublin mobility using multi-modal data. With longer datasets and real-time integration, the system could become a powerful tool for transport planning, forecasting, and public communication.

## REFERENCES

- [1] Smart Dublin / SDCC, “**SCATS Traffic Flow Data — January to June 2023**,” Dublin City Council Open Data Portal, Accessed: Dec. 5, 2025.  
Available: <https://data.smartdublin.ie/>
- [2] Met Éireann, “**Daily Weather Observations (DLY532) — Phoenix Park**,” Irish Meteorological Service, Accessed: Dec. 6, 2025.  
Available: <https://www.met.ie/climate/available-data>
- [3] National Transport Authority (NTA), “**GTFS-Realtime Trip Updates (v2)**,” Transport for Ireland Open Data, Accessed: Dec. 7, 2025.  
Available: <https://api.nationaltransport.ie/>
- [4] National Transport Authority (NTA), “**GTFS Static Feed (routes.txt, trips.txt, stop\_times.txt, stops.txt)**,” Transport for Ireland GTFS Repository, Accessed: Dec. 7, 2025.  
Available: <https://www.transportforireland.ie/transitFeeds/>
- [5] Dublin Bus / Transport for Ireland (TFI), “**Realtime Passenger Information — Service Alerts**,” Transport for Ireland Open Data Portal, Accessed: Dec. 7, 2025.  
Available: <https://www.transportforireland.ie/open-data/>
- [6] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2012.
- [7] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From data mining to knowledge discovery in databases,” *AI Magazine*, vol. 17, no. 3, pp. 37–54, 1996.
- [8] J. W. Tukey, “The future of data analysis,” *Ann. Math. Stat.*, vol. 33, no. 1, pp. 1–67, 1962.
- [9] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2009.
- [10] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2021.
- [11] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [12] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA, USA: Morgan Kaufmann, 1993.
- [13] Y. Freund and R. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, 1997.
- [14] T. Joachims, “Text categorization with Support Vector Machines,” in *Proc. 10th Eur. Conf. Mach. Learn.*, 1998, pp. 137–142.
- [15] C. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.
- [16] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O’Reilly Media, 2009.
- [17] M. Collins, “Discriminative training methods for hidden Markov models,” *ACL*, vol. 2002, pp. 1–8, 2002.
- [18] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
- [19] R. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd ed., OTexts, 2021.
- [20] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Boston, MA, USA: Springer, 1998.