Contents

The brief

Hello, this is Steven from Thameslink. I had the idea to use public social media tweets to improve our vehicles availability. I believe we can identify defect components in the vehicles based on the tweets. My engineers extracted some data for me. I would need experts who can analyse the data and prove my idea. Can you help me ?

Target from Client

Predict Defaults Using Data Mining Understanding & cleaning the data, preparing it for statistical model building, evaluation and identification of the model with the highest explanatory power.

What the client asks for?

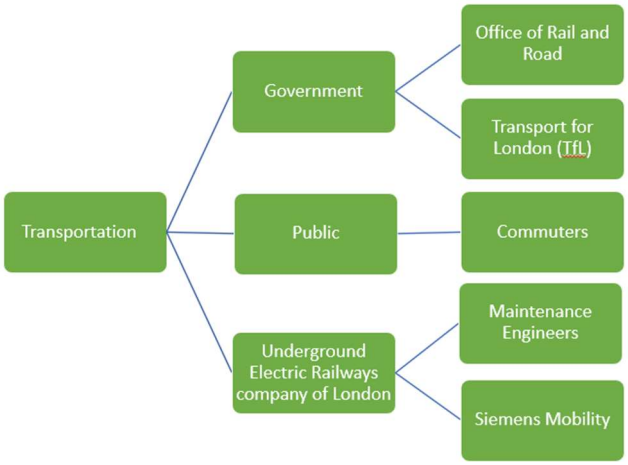
Improving London rail services based on tweets and sentiments of the users. Hence, incase there is a defect in the service provided, the product delivered should help fix the issues.

Business understanding

The London Underground (LU), popularly known as The Tube, spans nearly 400 route miles, with the busiest line of the system carrying over 180 million passengers. Logistics must cope with the occurrence of unpredictable events such as track or train failures, no-show of crews and so on. A robust solution is required for redressal of issues voiced by the customers and conflicts within the framework of the London Underground must be made.

Stakeholders

1. Underground Electric Railways company of London
2. Office of Rail and Road
3. Transport for London (TfL)
4. Commuters (Public)
5. Siemens Mobility



Project Goal

The project goal is to use data given by the client and Deliver a ML model which can classify sentiments based on tweets.

Most Viable Product:

The business objective is to provide better service to the customers of the London Underground and in turn reduce the negative sentiments posted by the customers on twitter. These Negative comments gave to be mitigated by solving the grievances of the users.

Thus, the MVP will be a Machine Learning model is "**to classify tweet sentiments**" thus:

- this model can help new tweets to analyse and classify sentiments.
- Once the model classifies sentiments of new text, we can do in-depth keyword-extraction from text.
- These extracted keywords can solve our business challenge of poor train service.

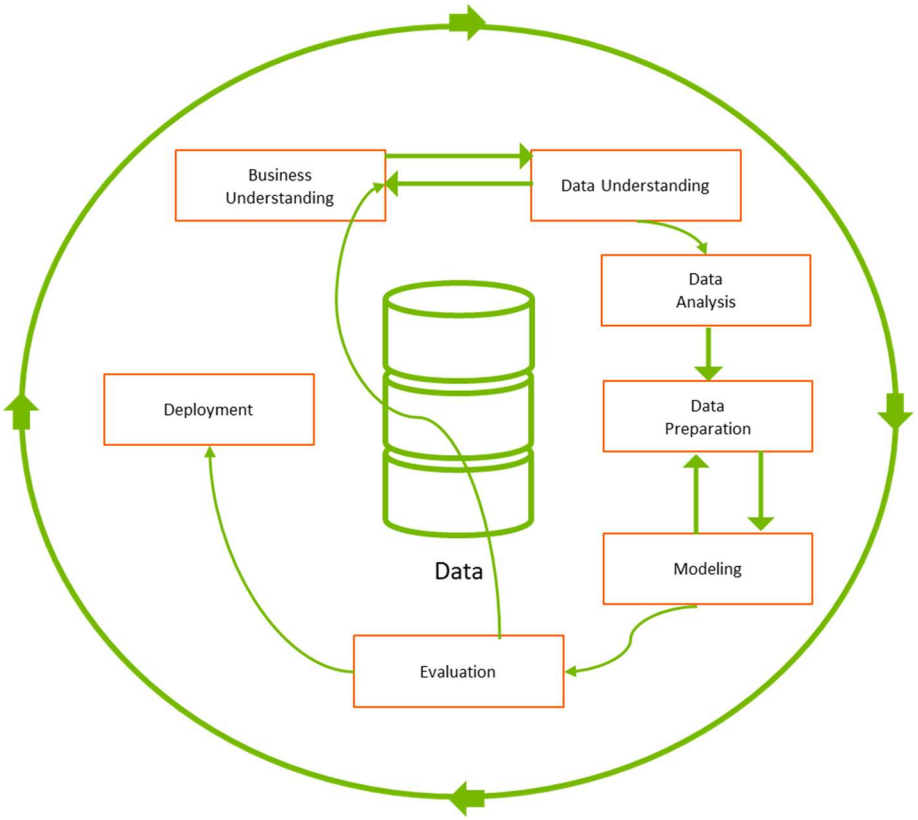
Project Roadmap based on Agile Scrum

Sprint 1	Sprint 2	Sprint 3	Sprint 4
• Business Understanding • Project Goals • Project Roadmap • Data Understanding • EDA -1	• EDA – 2 • Data Preparation and Cleaning • Feature Engineering • Modelling	• Model Deployment • Model Evaluation • Project Delivery	• Delivery of MVP • Client Presentation

Project tools and processes:

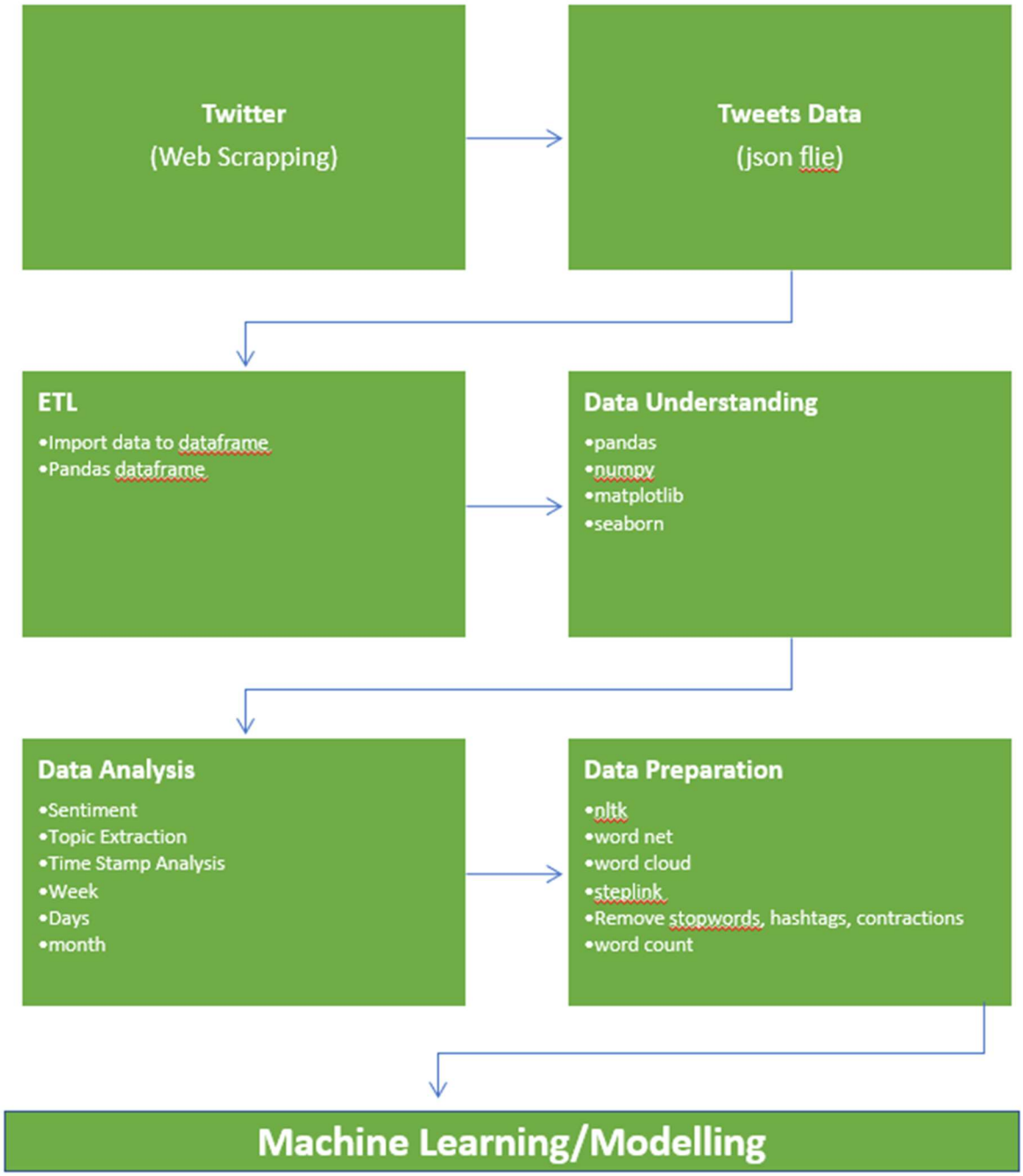
Processes	Technical Aspect	Tech Stack
• Agile Methodology: Scrum • Tool: Trello Board	• Data Science and Machine learning using CRISP Data Model	• Jupyter Notebook • Python • Data Science • Machine Learning • NLP

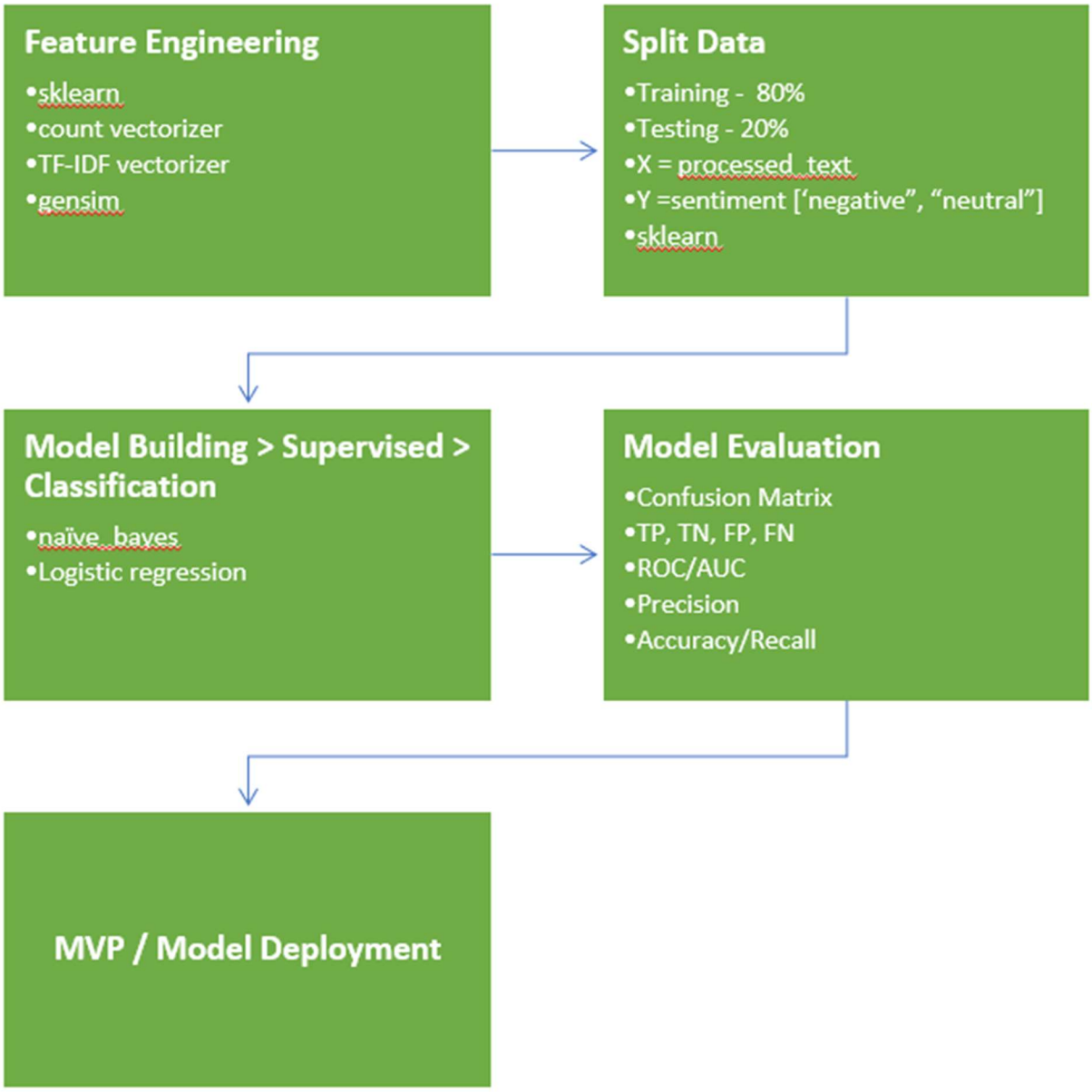
Crisp DM for Modelling:



For the technical aspect of the project, we will be using CRISP DM model to build and deliver the DS – ML model Architecture of the Model

Architecture




Feature Engineering
•sklearn
•count vectorizer
•TF-IDF vectorizer
•gensim
Split Data
•Training - 80%
•Testing - 20%
•X = processed_text
•Y =sentiment ['negative", "neutral"]
•sklearn
Model Building > Supervised > Classification
•naïve_bayes
•Logistic regression
Model Evaluation
•Confusion Matrix
•TP, TN, FP, FN
•ROC/AUC
•Precision
•Accuracy/Recall
MVP / Model Deployment

Data Understanding

Key information of the given data

1. Data is projection of negative sentiments via tweets by the daily transport service users especially passengers of the trains. The data projects negative, neutral and positive sentiments. Most of the tweets have negative sentiments attributed to them.

2. The data provides details about

- when the tweets were made
- who made the tweets (author id of the tweets)
- tweet itself (the text of the tweets)
- language of the tweets (English)
- time of the tweet – day month year, hour minute and second
- location of the tweeter (coordinates in latitude and longitude)
- source (sprinklr, brandwatch)

3. The label information provides the label of the tweets -here we get sentiments text and other such details of the data.

4. The tweets are labelled negative, positive, neutral which in order projects dissatisfaction, happiness or probably just a comment

5. There are 16949 tweets

6. The tweet data has topics implying which are the key topics the users are tweeting about for the transportation service which are like, delay, service, cancellations, places.

Data Preparation

1. The columns featuring date and time, have been further split into year, month, day, week and time hours. New columns called “year”, “month”, “day_of_week”, “is_weekend”, “hour”, “dayparts” have been added to the dataframe as shown in the figure below.

year	month	day_of_week	is_weekend	hour	dayparts
2020	9	4	0	21	evening
2020	10	1	0	7	morning
2020	10	0	0	19	evening
2020	10	0	0	19	evening
2020	9	0	0	11	noon

2. There is information related to the author id, and tweet id, which will be ignored as they do not add any value for the EDA

3. The longitude and latitude were good data points to understand the precise location and status of the person tweeting., this would have given us an insight into why the person is tweeting while being at a specific train halt. However, since 91% of the data under this column is missing, this column will not be considered for the EDA and modelling.

4. The label column is extracted to understand the critical features: topic and sentiment.

labels	source_id	sentiment	topic
{'topic': [{'tweet_id': 'acd7673f-e621-5f1a-d6...	NaN	negative	service
{'topic': [{'tweet_id': '5b92aba8-4b05-6c63-84...	NaN	negative	delays
{'topic': [{'tweet_id': '0a799c07-8b76-17ba-b8...	NaN	negative	toilets
{'topic': [{'tweet_id': '8b4d2a34-c4f0-0e19-40...	NaN	negative	toilets
{'topic': [{'tweet_id': '1fd08862-d8c7-0682-6b...	NaN	neutral	seats

5. The topic has the following words frequently used to express sentiments 'handrails', 'toilets', 'plugs', 'none', 'service', 'tickets/seat_reservations', 'noise', 'covid', 'vandalism', 'roof', 'floor', 'wifi', 'train_general', 'brakes', 'tables', 'station', 'announcements', 'delays', 'air conditioning', 'seats', 'doors', 'windows', 'hvac'. All these words appearing in the topic column have either of the three sentiments attributed to them. This gives us more insights into what the data suggests.

6. The Sentiment column, among all its instances has 3 values: negative, positive and neutral. All the tweets are attributed to one of these three sentiments.

7. The importance is to clean the text from the tweets to be able to process them for better analysis and modelling. Hence a new column will be created “processed_text” from “text”.

	text	processed_text
0	@DSisourath The Thameslink core between London...	thameslink core london st pancras london black...
1	@DulwichHistory Loving the complaint about peo...	love complaint wait minute train they clearly ...
2	@SW_Help .And yet you have no toilets on some ...	and toilet train _railway manage every single
3	@SW_Help you have no toilets on some of your t...	toilet train talk _railway manage every single
4	@SpeedySticks007 @MrNeilJH @TLRailUK @christia...	daft care money backside seat disable adaptati...

8. Removed special characters from the tweets such as #@%/;$()~_? so that the algorithm will have a clear data set to extract findings such as making a word cloud or finding frequently used words in the tweets and so on.

9. Removed web links from the tweets as the web links will not contain necessary words or data to find topics and sentiments to do our analyses.

10. Handling stop words such as “the”, “in”, “an” which do not add any value to the algorithm as they do not reflect any sentiments by themselves. The stop words are disregarded without as far as meaning of the sentence is not jeopardized.

11. Words having apostrophe are changed from contracted words to two different words by removing apostrophe (eg: "who're”: "who are")

12. Removed tags as the tags will not carry any significance in finding words associated with the sentiments.

13. Remove digits all the digits are removed as here, in this data we are analyzing only words hence the need of numbers from this data is negated.

14. Fix misspelled words so that the algorithm catches the relevant words and no word is left out.

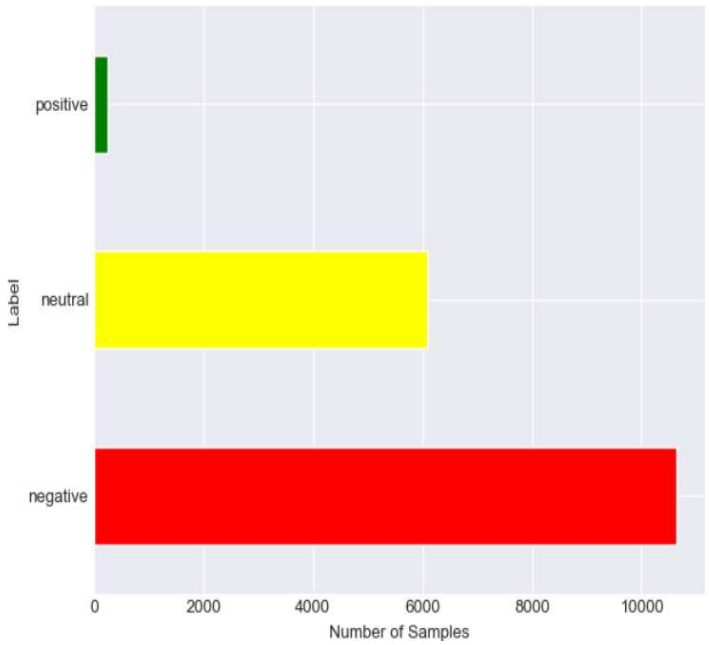
After making all the changes mentioned above, this is how the data from “processed_text” column looks like:

processed_text
thameslink core london st pancras london black...
love complaint wait minute train they clearly ...
and toilet train _railway manage every single
toilet train talk _railway manage every single
daft care money backside seat disable adaptati...
...
haha oh man audio corruption quite entertainin...
sweetis plug charge phone
far few commuter stand cram train put table ba...
vote thameslink pandemic period easier social ...
vote thameslink pandemic period easier social ...

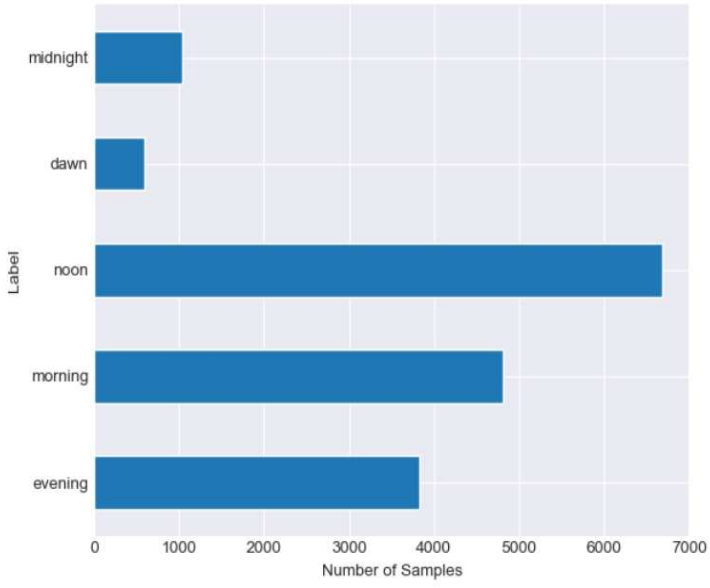
Based on the above observation we have derived various representations of our findings such as word count graph, word cloud whereby one can understand clearly the words that are associated with negative positive or neutral tweets.

```
0    thameslink core london st pancras london black...
1    love complaint wait minute train they clearly ...
2    and yet toilet train _railway manage every single
3       toilet train talk _railway manage every single
4    daft care money backside seat disable adaptati...
Name: processed_text, dtype: object
```

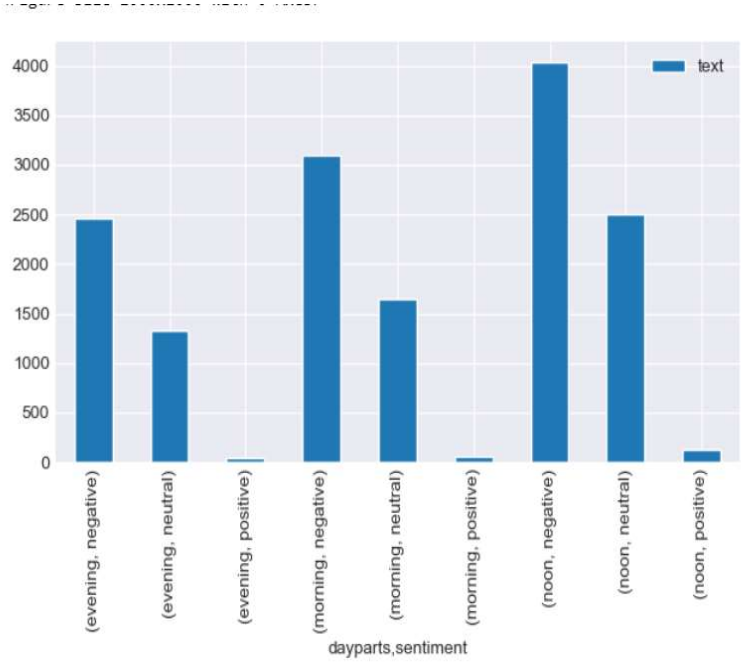
Data Analysis



Bar graph showing number of Negative, Neutral and Positive tweets



Frequency of Tweets at different times of the day

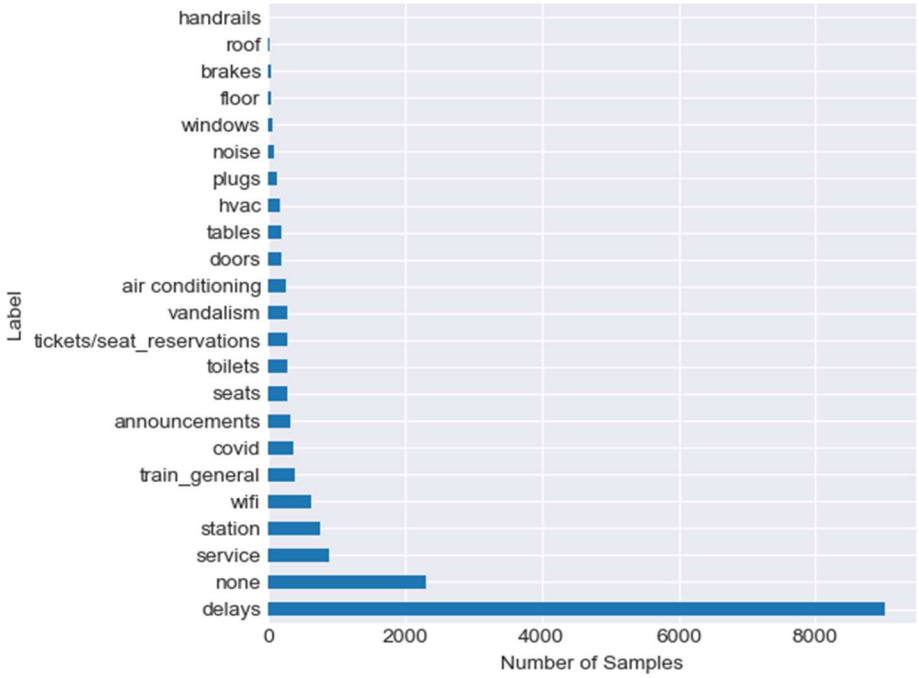


Comparison of Frequency of Positive, Negative and Neutral Tweets in different parts of the day

Summary:

There are more tweets appearing in the noon, that include negative, positive and negative sentiments most of the tweets are related to delays and service of London underground. So therefore, they have to look into the problems associated with delays and service of the train service. But the percentage rate per most of the tweets made are during the afternoon. It would be advisable deeply scrutinize why are these tweets about delays are occurring in the afternoon time.

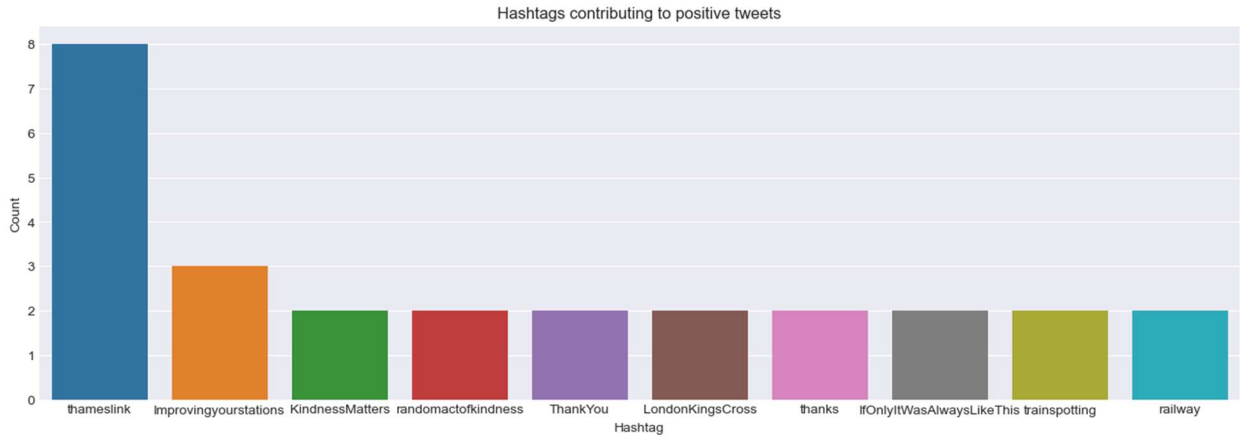
Most frequently appearing words in the data

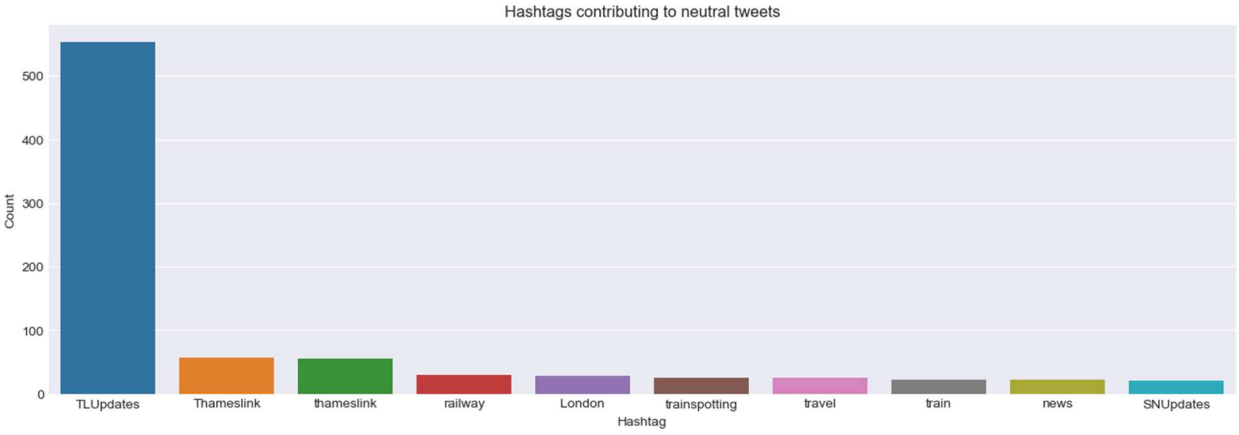


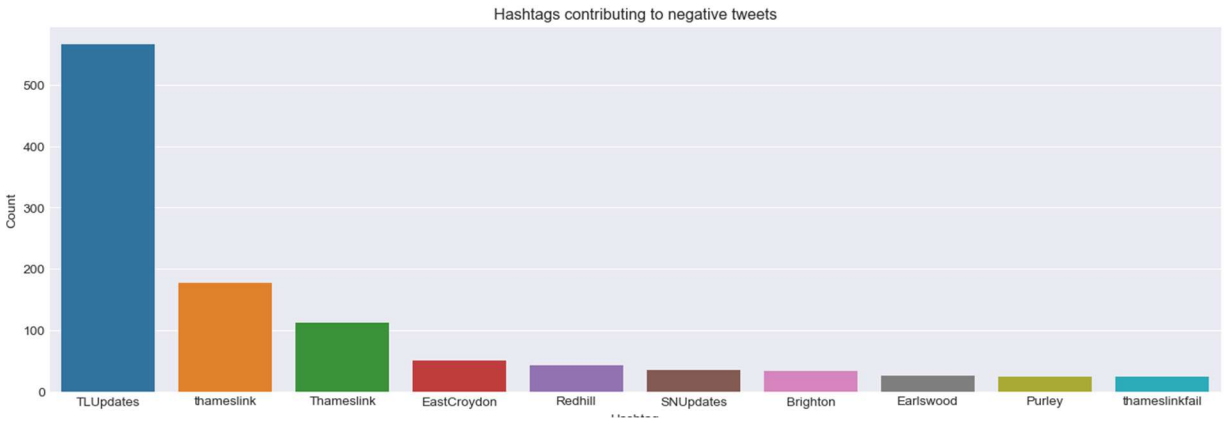
Topics of Tweets appearing in the Data

1. Delays – It seems the most tweeted negative comment is about delays, as mentioned before most of the tweets about delays are posted in the afternoon. One can deduce that the delays caused by the London underground are either in the afternoon or the people tweeting about the delays are posting it in the afternoon
2. None – None is a word caught by the model since it is a negative word, but one cannot determine what is the specific issue herein
3. Service – there is significant number of negative tweets about service in the data, so it is advisable that the client may take a look at the tweets to understand the issue raised by the passengers.
4. Station – The tweets about the station are quite a few. It could be about different as[ects of the service provided by the London Underground.
5. Wifi – few but significant number of tweets holding a negative sentiment about wifi have been posted by the passengers.
6. Train General, wifi, covid, toilets, seats and so on are self-explanatory

The end user can benefit from the extracted keywords, wordcount graph and hashtag analysis to arrive at root cause analysis so as to understand the reason behind negative tweets. Hashtags are precise keywords that point out what the person tweeting is pointing out at. Hence hashtag analysis should be considered an important tool to understand under which light has the passenger tweeted.

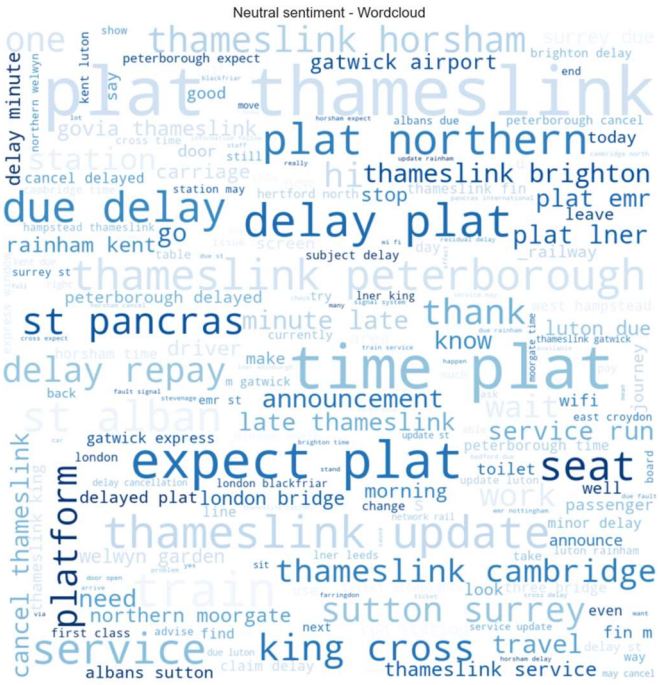




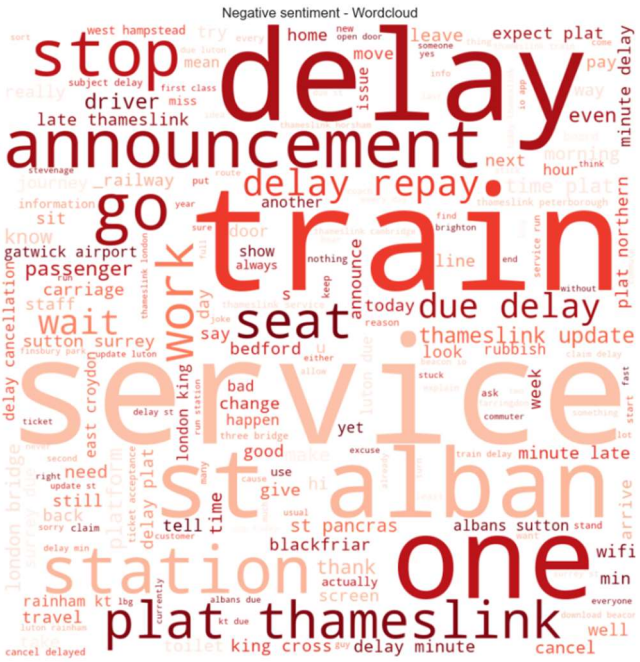


Below are word clouds which will give us a comprehensive idea of various words used by the passengers in the tweets:

Above is the word cloud that will give an idea of the words used and attributed to positive sentiments



above are the words appearing in the wordcloud associated with neutral sentiments



Above are the words appearing in the wordcloud associated with negative sentiments

Above are words appearing in bigrams in the given data. Bigrams are words a pair of consecutive written units such as letters, syllables, or words.

Modelling

- This model can help us stratify the data in different segments and analyze it and classify negative sentiments.
- Once model is set up with the data, sentiments attributed to the tweets can be extracted from the new text column called “processed _text”, we can do in-depth keyword-extraction from this text.
- These extracted keywords can solve our business challenge of poor train service.

Feature: processed text

Target: sentiments [negative, neutral] # Note dropping positive due to lower available data and avoiding imbalance dataset problem

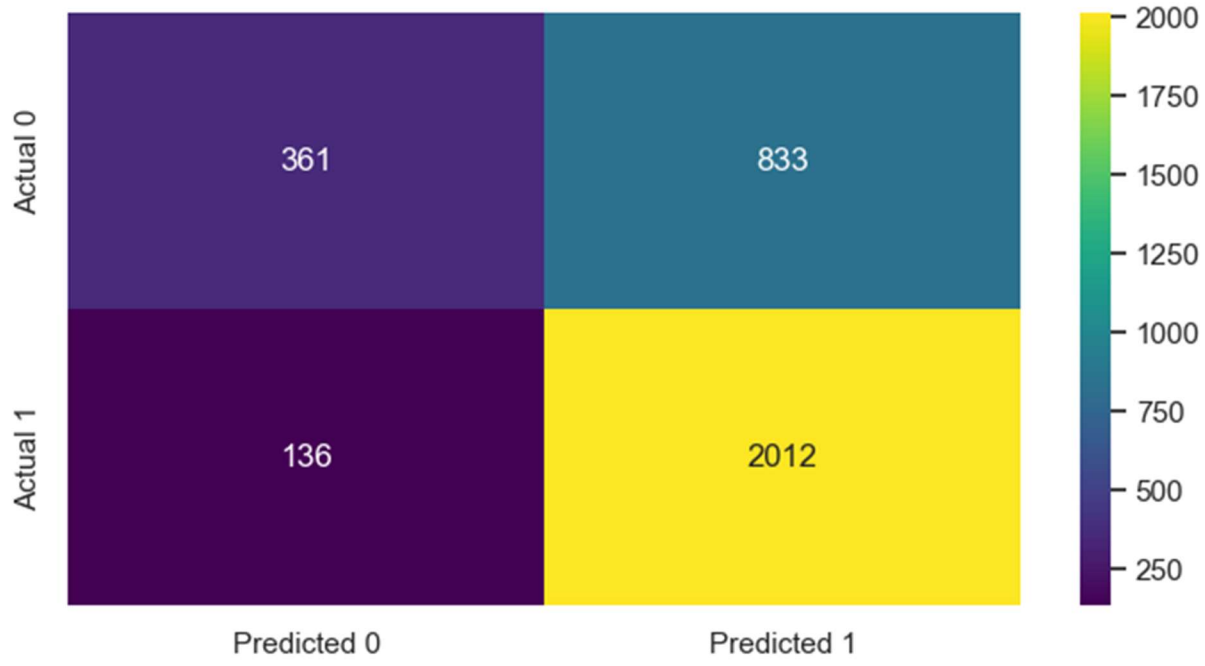
#For binary classification converting text into 1 and 0

```
target = {

   'negative': 1,

   'neutral': 0

}
```

The model is trained using the Naïve bayes classification model available from scikit.

Model Evaluation and Performance

Confusion Matrix



Total number of TPs: 2012

Total number of TNs: 361

Total number of FPs: 833

Total number of FNs: 136

	precision	recall	f1-score	support
0	0.73	0.30	0.43	1194
1	0.71	0.94	0.81	2148
accuracy			0.71	3342
macro avg	0.72	0.62	0.62	3342
weighted avg	0.71	0.71	0.67	3342

Performance Evaluation of the model

Important Performance Metrics In the Delivered Model

1.**Precision**: Precision is defined as the ratio of correctly classified positive samples (True Positive) to a total number of classified positive samples (either correctly or incorrectly).

2.**Recall**: We want to predict the number of negative tweets correctly. More the number of False negatives, (Negative but predicted as neutral) will decrease the necessary actions to be made to improve the service. Hence recall has to be increased and FNs should be decreased.

3. **F1-score**: Harmonic mean of Precesion and Recall. Precision is related to FPs. Precision and recall are connected. Reduce the false positives.

Confusion matrix provides a good result to understand the performance of the model. We have different measure criterion as options to choose from TP, FP, TN, FN, F1-SCORE, ACCURACY, RECALL, and PRECESION as explained above. However, Recall is the most suitable measuring criteria for this model.

Henceforth, after this evaluation it can be concluded that:

- The target vision of this project is to reduce the amount of negative tweets.
- Therefore, here is a model which include extracted keywords.
- Which leads to negative tweets arising from dissatisfaction in the users.
- The most important evaluation value is therefore the Recall

MVP delivered in this project

In this project a Machine Learning model has been delivered to the client. It is a machine learning model which can analyze text from tweets and segregate it to negative, neutral and positive. It extracts keywords mostly on topics that help them gather information which is key in the process of estimating the tweets that contribute to negative or neutral and thus one can take decisions so as to understand in which areas of service needs improvement.

Thus, the specific stakeholders then can work on the relevant topics which can help them address the specific issues.

Below progress of the project can be seen at the completion of the project