# Sentiment Analysis for Public Outreach regarding Transportation Projects using Natural Language Processing

## Problem Statement and proposed Solution

Pena Boulevard (a long stretch of freeway from Downtown Denver to the Denver Airport is the key to connecting the city of Denver to places around the country (dare we say, world). According to the recent bill by the congresswomen, the workforce at the airport and the demand for the DEN airport will double in the next 20-odd years. Several stakeholders have proposed steps to cope with the increase in demand. With Environment compliance being a significant factor as well, we must keep the increase in vehicular pollution in mind to widen the roadways incessantly. Hence, we need a way to look at what the public thinks about future transportation endeavors. A tool that looks at conflicting ideas and filters out important information would be handy for stakeholders and transportation practitioners to manage their future objectives better.


**Figure 1: Word Cloud for Public Sentiments**

Public Outreach is an important step in most transportation projects as that drives most of the directions in the decade long projects. The public outreach goes in multiple rounds. Having a model with transportation related context would require some initial legwork to analyze, with all the manual tagging and natural language processing, but once we have a reliable sentiment analysis tool, we can use it to use it to drive decisions for the upcoming rounds of public outreach without having to do manual work.

**Table 1: Sentiment Tag Summary**

| Index | Tag | Count |
|---|---|---|
| 0 | A-Line: more reliable, higher frequency, faster or express A-Line service to DEN | 165 |
| 1 | Transit: direct transit services from metro Denver cities/suburbs to DEN | 89 |
| 2 | Transit: Cheaper A-Line and/or bus fares to DEN | 84 |
| 3 | Transit: More reliable and/or higher frequency transit service | 72 |
| 4 | Transit: Better connections to A-Line and/or between RTD services | 46 |
| 5 | Bike Infrastructure: Better bike facilities and/or bike connections to transit | 43 |
| 6 | Transit: Safer and/or cleaner transit vehicles and/or stations | 35 |
| 7 | Transit: More off-peak transit service | 33 |
| 8 | Park & ride: Cheaper or more convenient park & ride facilities with ample parkings and services to DEN | 33 |
| 9 | Transit: more convenient / easier access to transit | 30 |
| 10 | Transit: Dedicated bus lanes, better bus services or bus rapid transit (BRT) | 27 |
| 11 | Park & ride: safer and/or more secure park & ride facilities and/or covered parking | 23 |
| 12 | Transit: Already use and/or supports the use of sustainable transport modes | 22 |
| 13 | A-Line: Additional stops and/or add 72nd and Himalaya Station | 12 |
| 14 | A-Line: double tracking to reduce delays and increase frequency | 11 |
| 15 | Bike Infrastructure: Secure bike storage at RTD stations | 11 |
| 16 | Park & ride: p&r facility at I-70 & Pena with connection to high speed transit | 7 |
| 17 | Ped Safety: Safer pedestrian facilities and/or complete streets | 7 |
| 18 | Transit: Promote and incentivize public transportation, provide more information to the public on the RTD system and/or transfers to the A-Line | 5 |
| 19 | Park & ride: guaranteed and cheaper parking at 40th & Airport station in addition to convenient access from Pena. | 4 |
| 20 | Suggestion for improvements outside of Study area | 3 |
| 21 | Personal Rapid Transit (PRT), Employee shuttles and free eco-pass for employees | 3 |
| 22 | Transit: Additional racks/space for luggage on transit | 2 |
| 23 | No Tag | 162 |

## Workflow

Facebooks *fastText* embedding (**Figure 2**) is used to embed the sentiments and convert them to vectors representing them. Also, the initial rounds of sentiments had to be manually tagged using engineering judgement. Once the manual tagging is done, individual binary classifiers were built for individual tags to predict the key takeaways for each comment. Multiple classifying algorithms were used for all the binary classifiers and the best algorithm for each sentiment was chosen.

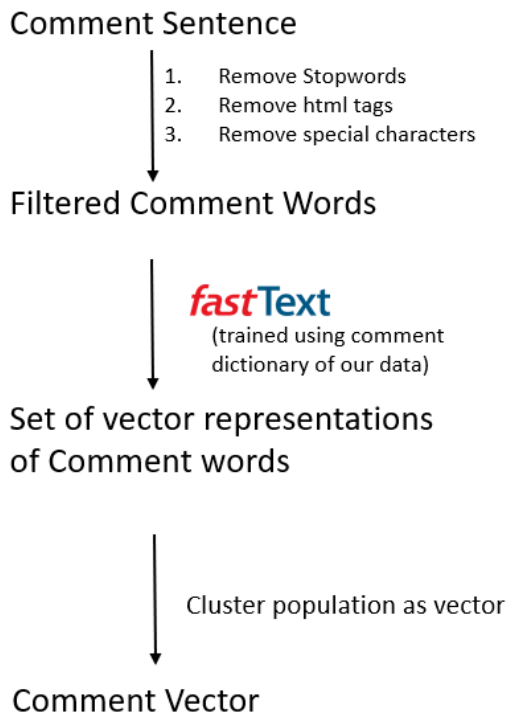


**Figure 2: Embedding vectors for Public Comments**

## Multiple ways of using *fastText* embedding

The following methods (Figure 3) were used to convert comments to vectors and the best performing method was used moving forward.


Figure 3: Two methods of using fastText embedding

This did not provide enough success in the beginning and hence a new strategy was implemented. Firstly, a mode choice classifier was trained and the beyond that each individual sentiment related to the sentiment was trained.

## Baseline

To compare the performance of the model, a baseline model was used. In case. The average vector per sentiment was computed and the comments that have the lowest cosine similarity with the average vector of a said tag was associated with the tag. Going by **Figure 4**, the baseline does not do well for most sentiments.

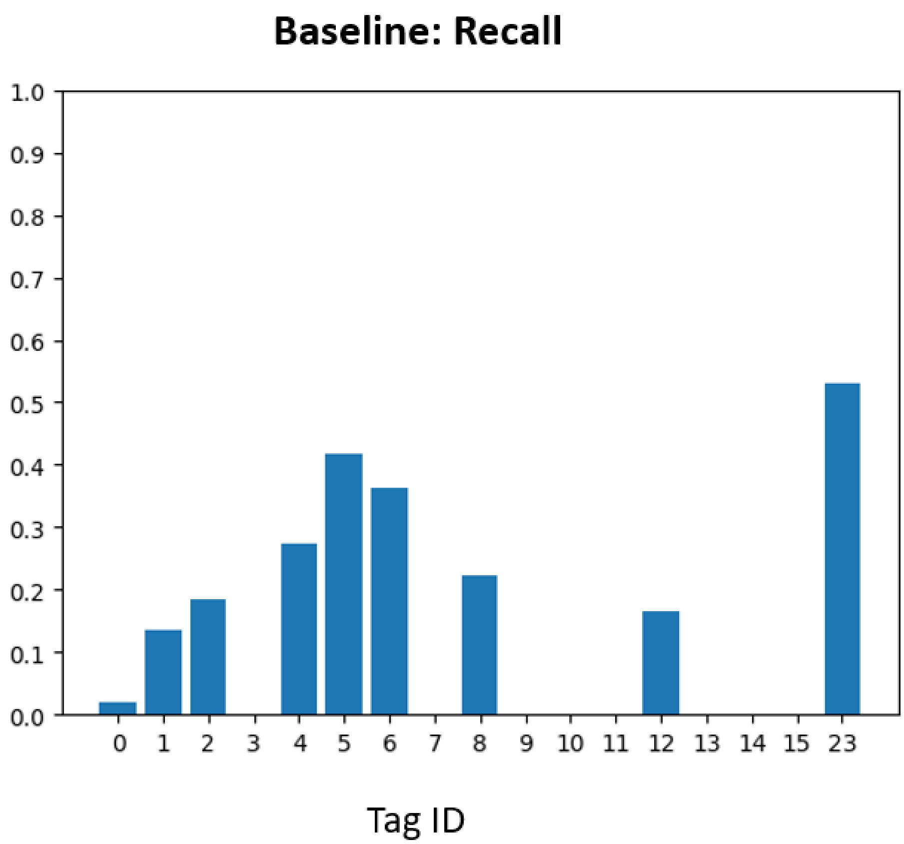


**Figure 4: Baseline: Recall values for each sentiment tag**

## Mode Choice based NLP model.

As a case study, transit model choice for sentiment analysis is showcased. As indicated by **Figure 5**, multiple classifying algorithms were trained and in the end a mixture of the best binary classifiers were used to classify each mode-choice based on each comment. On top of the transit mode choice, classifier models for individual sentiment related to transit were trained and a mixture of the best transit related sentiment predictors were chosen as the best predictors of individual sentiments.

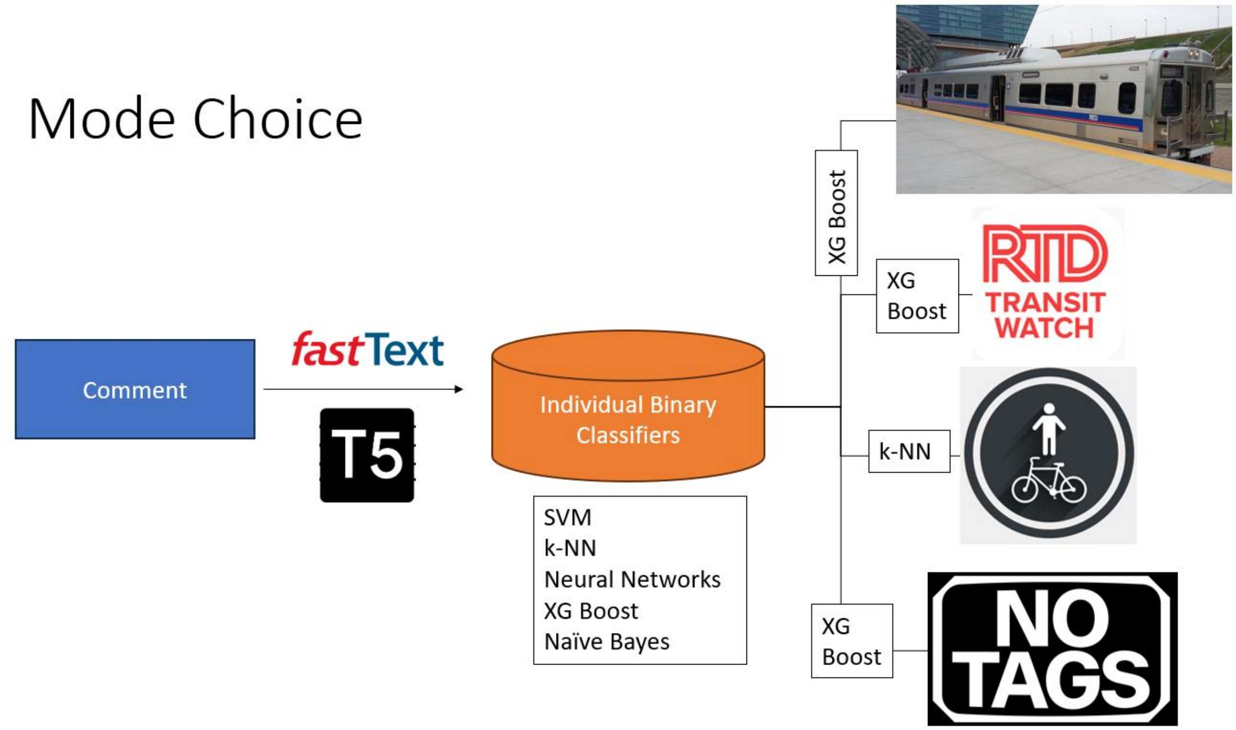


**Figure 5: Mode Choice Methodology for Primary Tags**

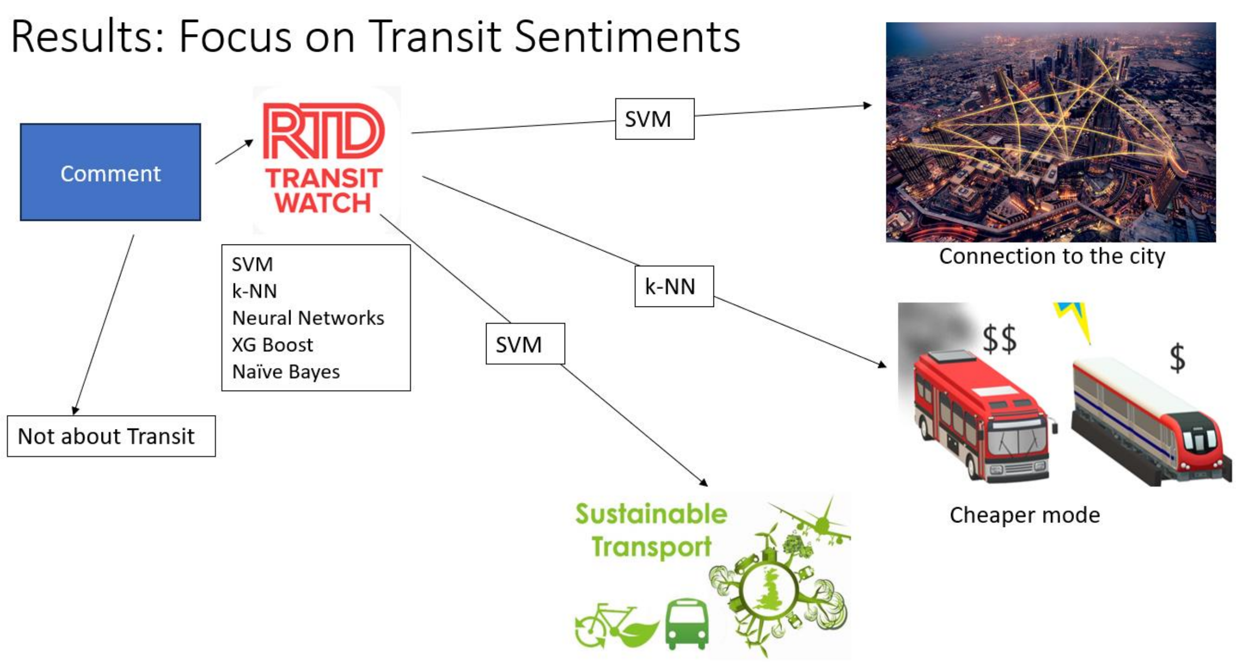


**Figure 6: Transit Sentiments Classifying Methodology**

## Results

As indicated by **Figure 7**, we see significant improvement compared to what we observed with our baseline. However, some sentiments did not perform that well. Upon further investigation, it was noticed that comments related to sentiments like "connection to the city" were more wordy and less concise making it harder to interpret it without human help. Having more population under each sentiment would possibly make out models more robust.

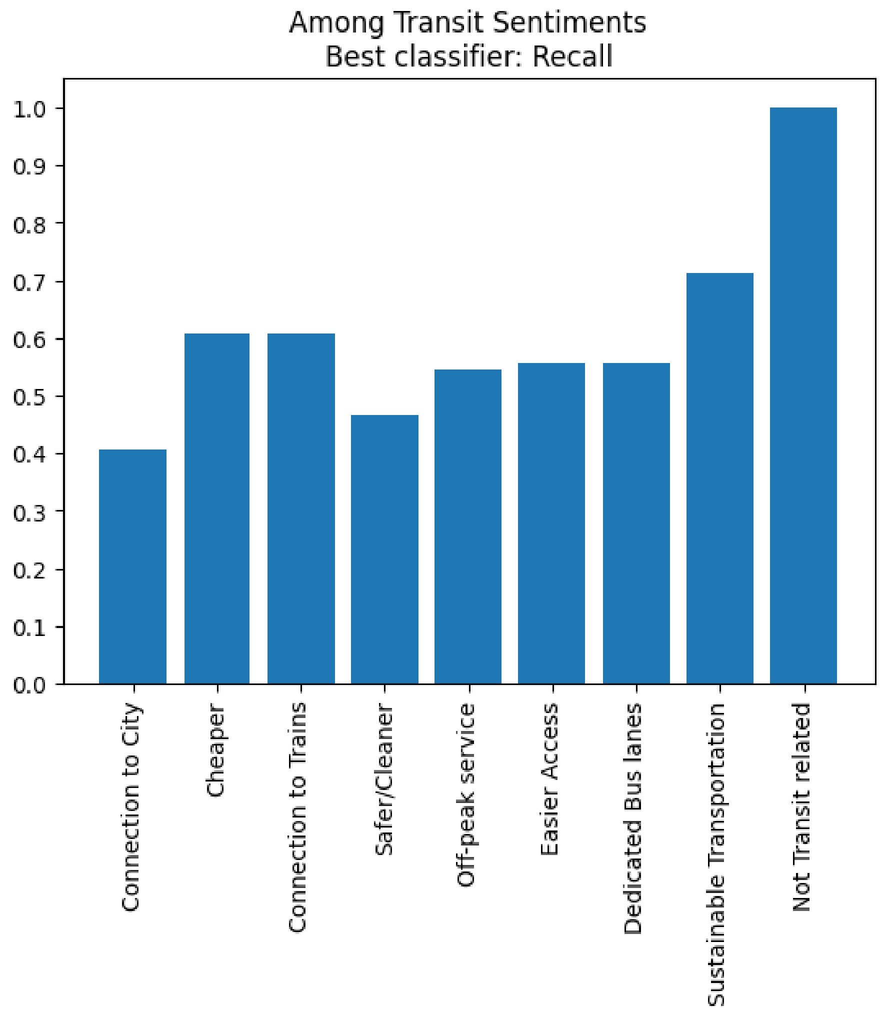


**Figure 7: Transit related Classifier Performance**

### Limitations and Future Work

- A lot of comments were unclear about the needs for the commenters.
    - More directed survey would probably make the exercise more informative.
- Using LLMs to get executive summaries may suffice compared to the needs of the project.
- Having more rounds of engagement will give more data to train models with, making the models more robust.