CE 597 Project proposal

Title: Analyze User Experience and Safety in the NYC Subway System using Natural Language Processing (NLP)

Group (Rollin' Rollercoaster): Adway Das, Abhishek kumar Prajapati, Pengxiang Zhang

Introduction

Public transportation plays a crucial role in urban life as an affordable and eco-friendly mode of transportation for individuals. However, the safety and dependability of these transportation systems have long been a concern for users, particularly in high-density cities like New York City where the subway system is one of the largest and busiest in the world. However, the safety of the NYC subway system has been a major concern for riders and the public for many years. Crime, such as theft and assault, continues to be an issue in the subway system (NYPD, 2021). Additionally, equipment malfunctions, such as signal and track problems, have led to delays and disruptions in service (MTA, 2021). Furthermore, issues such as overcrowding and inadequate access to emergency exits raise concerns about the preparedness of the system in emergencies (Komanoff, 2020). Understanding the perceptions and experiences of NYC subway users is essential for enhancing its reliability and safety, which is where NLP techniques can be useful (Pinheiro et. Al., 2010). With NLP, we can analyze vast amounts of user-generated data, such as tweets, to gain a more comprehensive understanding of the safety and reliability (Al-Sahar, 2021) of the NYC subway system, as well as the experiences and perceptions of its users.

Objectives

The main objective of this project is to use Natural Language Processing (NLP) techniques to analyze tweets related to the NYC subway system and gain insights into the user experience and safety of the subway. The specific objectives are as follows:

[1] To collect and pre-process a large dataset of tweets related to the NYC subway system.

[2] To perform sentiment analysis on the tweets to understand the overall sentiment of users towards the subway system.

[3] To identify and extract key topics and themes mentioned in the tweets related to the user experience and safety of the subway.

[4] To visualize the results and draw insights into the user experience and safety of the subway system based on the tweet analysis.

Methods

Data Collection: The data will be collected using the Twitter API. The API will be used to search for tweets containing keywords, hashtags, and mentions related to the NYC subway system.

Data Preprocessing: The gathered data will undergo preprocessing to eliminate any irrelevant information, such as URLs and mentions, and to rectify spelling and grammatical mistakes. Afterward, some keywords collected from a subset of the data will be categorized into distinct labels manually based on the expressions of users regarding the safety of the NYC subway system.

Sentiment Analysis: The processed data will be subjected to sentiment analysis to determine the overall sentiment expressed in the tweets. This will provide insights into the users' experiences with the subway system.

Safety Analysis: NLP techniques will be used to identify and classify specific safety-related issues mentioned in the tweets, such as delays, breakdowns, and accidents. This will provide a comprehensive understanding of the safety of the NYC subway system. Following are the NLP based techniques which will be explored.

[1] **Named Entity Recognition (NER):** NER can be used to identify and extract specific entities in the tweets such as locations, organizations, and events. This information can then be used to identify tweets that mention specific safety-related issues, such as delays at a particular subway station.

[2] **Text Classification:** Text classification techniques, such as supervised learning algorithms (e.g., Support Vector Machines (SVM), Naive Bayes, etc.), can be used to classify the tweets into different safety-related categories, such as delays, breakdowns, and accidents.

[3] **Word Embeddings:** Word embeddings, such as Word2Vec, can be used to represent the tweets as numerical vectors. These vectors can then be used as inputs to machine learning algorithms to classify the tweets into different categories based on their meaning.

[4] **Rule-Based Systems:** Rule-based systems, such as regular expressions, can be used to identify specific keywords or patterns in the tweets that are indicative of safety-related issues, such as "delay" or "accident".

Topic Modeling (Optional): Topic modeling techniques, such as Latent Dirichlet Allocation (LDA), will be used to identify the most important topics discussed in the tweets related to the NYC subway system. This will provide insights into the areas of concern for the users and help prioritize improvements.

Visualization: The results of the sentiment analysis and topic modeling will be visualized using data visualization tools such as Matplotlib and Seaborn.

Expected Outcomes

The expected outcomes of this project are as follows:

[1] A dataset of pre-processed tweets related to the NYC subway system.

[2] Insights into the overall sentiment of users towards the subway system.

[3] A comprehensive understanding of the key topics and themes related to the user experience and safety of the subway system.

[4] Visual representations of the results that provide a clear understanding of the user experience and safety of the subway system.

Conclusion

This project aims to use NLP techniques to analyze tweets related to the NYC subway system and gain insights into the user experience and safety of the subway. The results of this project will be valuable for decision-makers and stakeholders in the public transportation sector and can be used to improve the subway system and provide a better experience for riders.

References

[1] Pinheiro, V., Furtado, V., Pequeno, T., & Nogueira, D. (2010, May). *Natural language processing based on semantic inferentialism for extracting crime information from text*. In 2010 IEEE International Conference on Intelligence and Security Informatics (pp. 19-24). IEEE.

[2] Komanoff, B. (2020). *Improving subway safety in New York City*. Transportation Alternatives. https://transalt.org/issues/subway-safety

[3] MTA. (2021). *Safety & security. Metropolitan Transportation Authority*. https://new.mta.info/safety-security

[4] NYPD. (2021). *Subway crime statistics. New York City Police Department*. https://www1.nyc.gov/site/nypd/stats/reports-analysis/subway-crime-stats.page

[5] NYC Transit Riders Council. (2021). Safety & security. https://www.ridertc.org/issues/safety-security/

[6] Al-Sahar, R. (2021). *Evaluating the Use of Twitter in Gauging the Effects of a Transit Service Intervention on Customer Satisfaction* (Doctoral dissertation, University of Toronto (Canada)).