New York City Street Cleanliness: Applying Text Mining Techniques to Social Media Information

Huijue Kelly Duan
Rutgers, The State University of New Jersey

Mauricio Codesso
Northeastern University

Zamil Alzamil
Majmaah University, Saudi Arabia

Objective

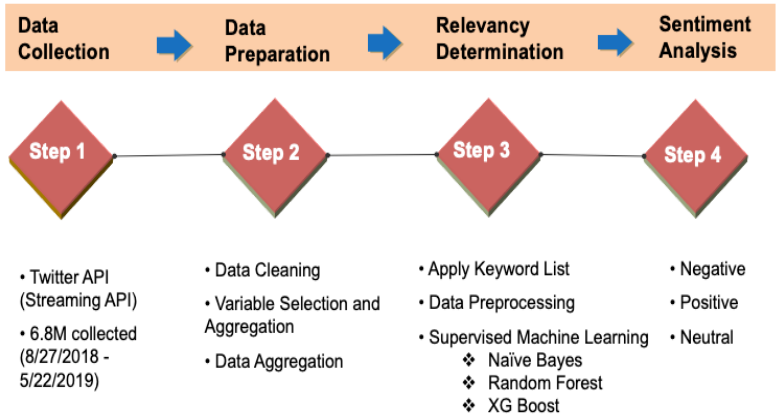
This study investigates the use of social media information and presents a framework for using crowdsourcing to evaluate people's opinion on NYC street cleanliness. It examines social media information by utilizing text mining techniques and different machine learning algorithms to identify the relevant tweets and assess the sentiment expressed in the content. The objective of this study is to:

- Identify temporal trends and patterns of the cleanliness of NYC street
- Analyze whether crowdsourcing information is consistent with NYC cleanliness ratings
- Assess the performance of municipal services via sentiment analysis

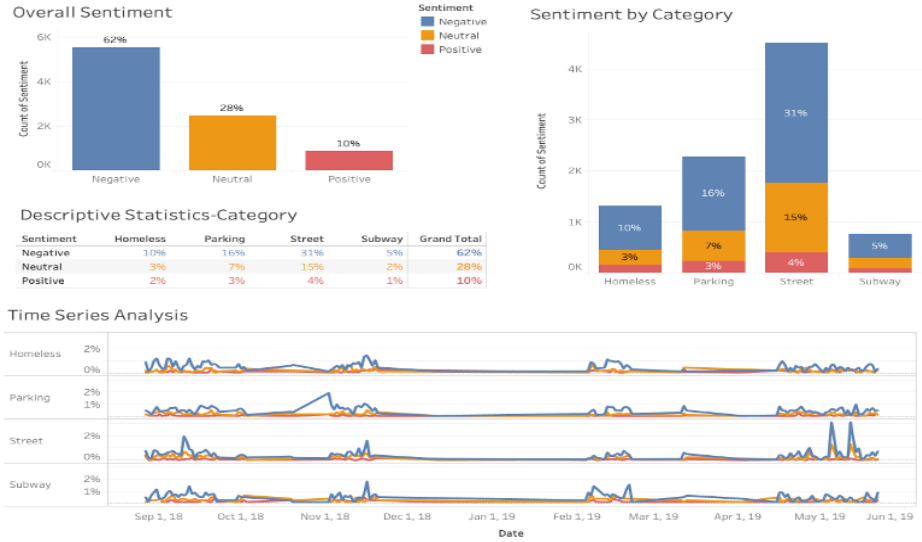
Relevancy Determination

- Apply a list of keywords to filter out some irrelevant tweets
- Preprocess the data by applying tokenization and lemmatization, removing hashtags, @, URL links, stopwords, etc.
- Use supervised machine learning models to identify relevant tweets, including Naïve Bayes, Random Forest, XG Boost
  - The dataset is facing an imbalanced classification issue, two sampling methods are used to solve the issue: Random under-sampling & Random over-sampling
  - Stratified 10-fold cross-validation with a paired t-test are performed to evaluate the performance of all classifiers (based on Accuracy, Precision, Recall, F-1 Score, ROC_AUC)

Workflow



Result



Conclusions

- The overall sentiment of the tweets is negative
- Majority of the negative tweets is related to street condition
- The framework is extended to another social media platform, Facebook. The incremental value of Twitter and Facebook is different, Twitter indicates more valuable information in this study
- Municipalities can continuously monitor the social media information, gain insight into people's opinions about municipal services, and to understand the residents' needs and complaints