City of Boston Priorities Media Sentiment Analysis

Yancheng Liu, Haiqi Ma, Karantonis Georgios, Ganz Faiz, Lu Peiqing
BU Spark! Hariri Institute of Computing, Boston University

Project Summary

The objective of this project is to understand the media coverage (news and social media) and the public response to the to the legislative agenda of Martin J. Walsh, the 54th Mayor of Boston. This implies understanding which of the agenda's priorities were covered by mainstream media, in what quantity, the time of response of mainstream media when covering them with respect to the press release, and the public's general sentiment regarding such topics. In order to achieve these goals, the priorities of the mayor's agenda have been identified through the City of Boston online database of press releases and a comparative statistical analysis of the coverage of these topics by mainstream media was performed in combination with a sentiment analysis of the public's opinion. The analysis performed was able to reveal the main topics that are covered by the mayor's agenda, the amount of coverage each of these topics received by mainstream media, the most discussed topics on social media, the change in coverage over time of mainstream media and the sentiment for each topic expressed by both mainstream and social media.

Methodology

Data Pre-Processing

- Data Filtering (FFNN)
- Removing stop-words, invalid characters, and digits (NTLK)
- Stemming and Lemmatization (NTLK)
- Feature Extraction (TF-IDF and doc2vec)

Priorities Identification

- Clustering , SVM, Cosine Similarity, and FFNN (Unsuccesfull)
- Latent Dirichlet Allocation (LDA)

Sentiment Analysis

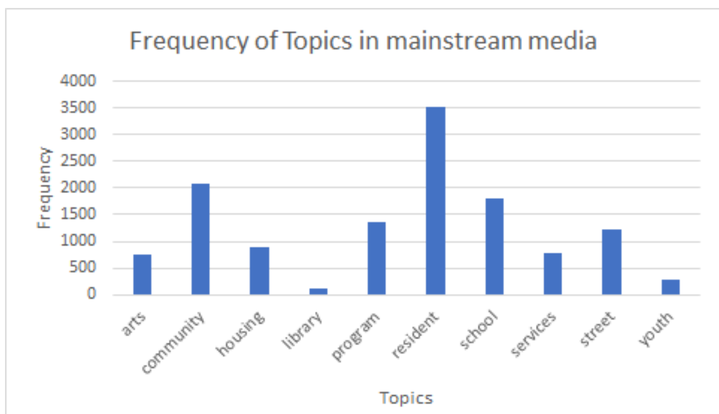
- TextBlob and Vader
- Long short-term memory (LSTM)
- Gluon NLP

Statistical Analysis

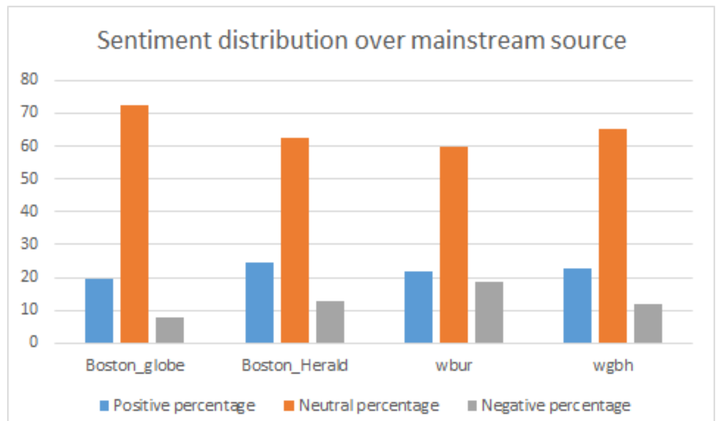
- Frequency Count (Coverage Quantity)
- Sentiment Distributions (Topics and Sources)
- Time Distributions (Mainstream Media and Topics)

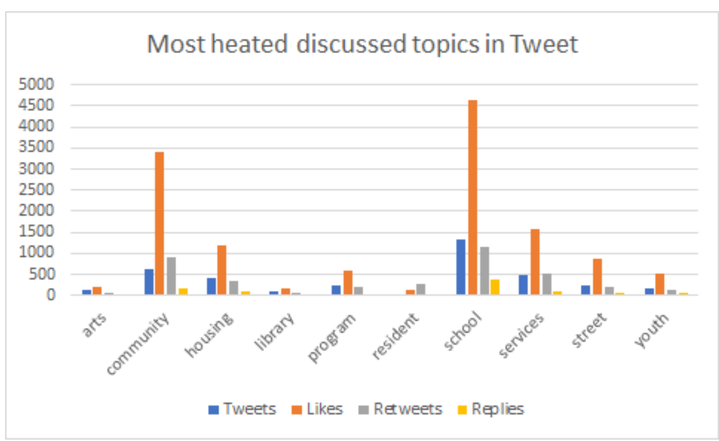
Results

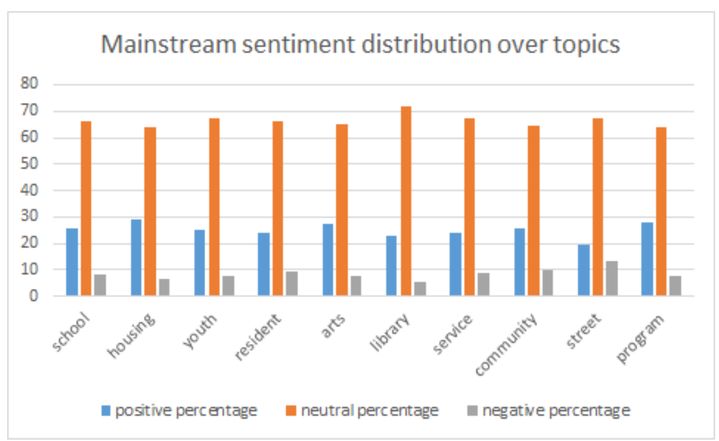
Priorities Coverage



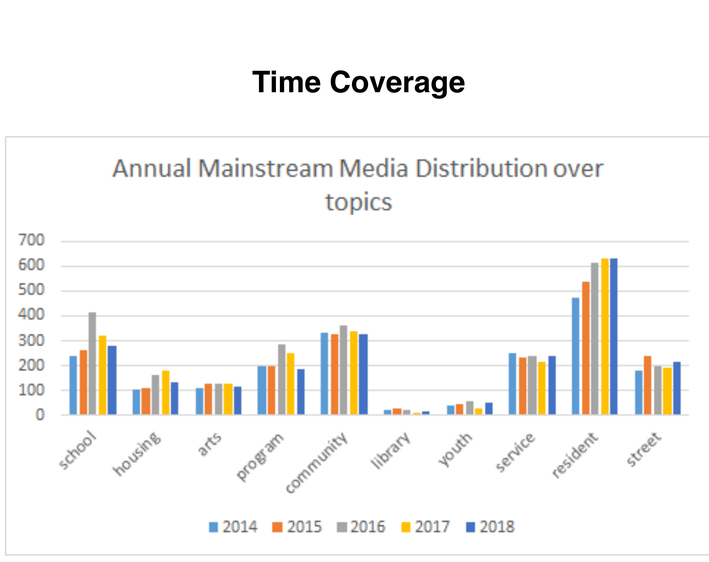
Sentiment Analysis Statistics

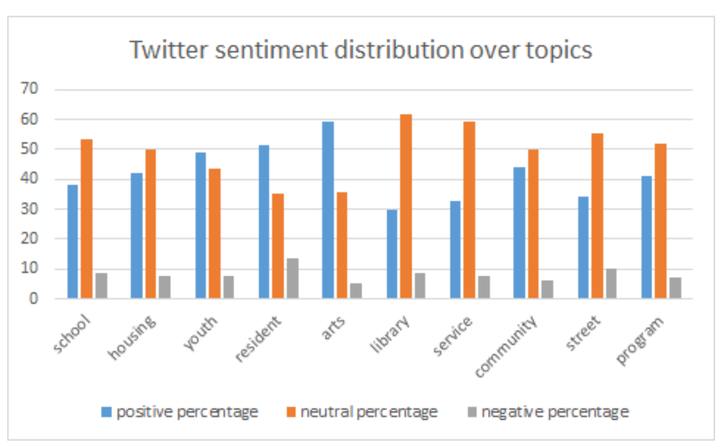






Time Coverage





Future Work

In this project, various methods to analyze the data from press release, mainstream media and social media were examined. The work performed could be further improved by collecting more labelled data fine tuning two of the system's submodules chosen in this project: the priorities' identification and the sentiment analysis.

Regarding the priorities' identification, the need for a concrete classification scheme is imperative. We believe that classification based on the documents' could give better results but additional tests and fine-tuning are required. Regarding the sentiment analysis engine, even though our best working model manages to achieve a high accuracy of 86%, it could be modified to score up to 93% by implementing the current state of the art model.

Finally, one of the biggest constraints we had to overcome was the lack of labelled data. We argue that finding sources, such as websites, from which such data could be extracted, could not only give a significant boost to our system's performance, but could also open new possibilities in the models that could be utilized.

Conclusion

For the purposes of this project, we created a collection of data by scraping the following four online media outlets; Boston Press Releases, The Boston Globe, The Boston Herald, WBUR, and WGBH, along with data collected using the Twitter API. The collected data expands from 2014 to this day and contains all the tweets and the articles, along with their metadata, detected in this time period. The topics of the mayor's agenda were determined by applying Latent Dirichlet Allocation on the City of Boston Press Releases dataset.

The mainstream media filtering was performed by training a 5 layer feed forward neural network on a sample of the original dataset, achieving an accuracy of 99.83%. The filtered articles were assigned to their corresponding agenda topics using a majority voting based classifier and finally, the classification results were passed to the Gluon NLP sentiment analysis system, that achieved an 86% accuracy.

The data point collected from Twitter were filtered using a keyword matching technique and were assigned to their corresponding agenda topics using a majority voting based classifier. The sentiment analysis was performed by a 2 layer LSTM network, that achieved an accuracy of 79.49%.

Finally, the same statistical analysis procedure was performed on all of the data allowing us to answer the majority of the priority questions.

Acknowledgments

We gratefully thank Lance Galletti for his support and mentorship and to the entirety of the BU Spark team for providing us with this insightful project which was a unique experience.