Predicting Upvotes and Replies on New York Times Article Comments

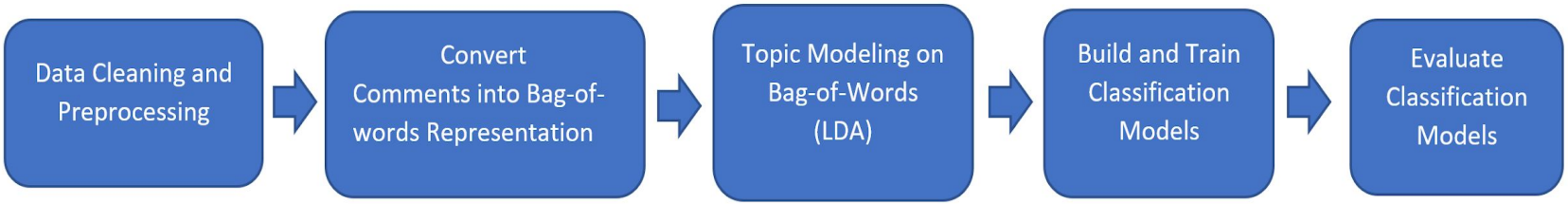
Abstract

- The New York Times is one of the major news outlets in the U.S., and the comment section gives information about many readers' opinions
- Number of replies and upvotes on these comments is a reflection of how people respond to certain opinions or what kinds of sentiments are most popular or controversial among readers
- We reduce a bigram bag of words for our dataset of comments to 18 LDA topics
- We then use these topics to predict number of upvotes and replies using linear regression, ridge regression, random forest regression, Naive Bayes, and logistic regression

Background

- Data: 231450 comments posted in January 2017 on a variety of articles
- New York Times in need of better way to filter through comments (currently has human moderators)
- Previous work classified comments with <4 upvotes as not popular and >4 upvotes as popular and examined correlation between popularity and certain features of comments
- Previous work on text analysis found certain punctuation marks to be useful for sentiment analysis
- Hierarchical Dirichlet Processes can be used to determine number of topics

Approach


Data Cleaning and Preprocessing
Convert Comments into Bag-of-words Representation
Topic Modeling on Bag-of-Words (LDA)
Build and Train Classification Models
Evaluate Classification Models

Data Pre-Processing

First, we cleaned the text data by removing HTML formatting and ignoring non-ASCII characters. Then, we edited the preprocessSentences.py script provided in Assignment 1 to tokenize the “?” and “!” punctuation characters and to process bigrams as well. Running this script on the comments, using only the tokens that appeared 100 or more times, converted the text data of the comments to a bag-of-words representation with 11493 features. These features included both unigrams and bigrams.

The bigram bag-of-words representation was analyzed with unsupervised learning through Latent Dirichlet Allocation. The optimal number of topics was determined by a Hierarchical Dirichlet Process to be about 18 topics, so we then initialized a Latent Dirichlet Allocation model to reduce the dimensionality of the text data to 18 features. We then examined the important words of each topic.

Topic Modelling

Topic Rank	Sample of Important Words
1	trump, president, new, go, time, !
2	trump, news, medium, lie, nyt, fake news
3	?, get, law, ? ?, order, congress
4	like, one, look, think, see, read, article
5	school, public, education, would, use, time
6	republican, democrat, senate, supreme court, nominee
7	trump, country, tax, return, ban, muslim
8	war, american, country, military, would, israel

Analyses

- Regression
  - Linear
  - Ridge
  - Random Forest
- Classification
  - Naive-Bayes
  - Logistic Regression

Constraints

- Number of replies are integers
- Most of the numbers are zero, very few are greater than ten.
- Number of replies varies with wide range

Improvements

- Use Classifiers
- Define the outcomes as ranges, not values

Performance of Regression Models Using Raw, Binary, and Range (5 categories) Data

Model	Raw R-squared	Raw MSE	Binary R-squared	Binary MSE	Range R-squared	Range MSE
Ridge	0.001121	6.149	0.00145	0.0238	-0.0101	0.0504
Lasso	-3.910e-5	6.156	-2.284e-6	0.0238	-0.0100	0.0503
Linear	0.001120	6.153	0.00145	0.0238	-0.0101	0.0504
Random Forest	-0.3106	8.068	-0.140	0.0271	-0.2096	0.06030

Table 1: Performance of classifiers for number of replies

Performance for upvotes

Model	Raw R-squared	Raw MSE	Binary R-squared	Binary MSE	Range R-squared	Range MSE
Ridge	0.003585	15202.81	0.0134	0.246	0.00341	2.304
Lasso	-6.154e-7	15257.513	-4.676e-7	0.249	-0.00787	2.330
Linear	0.00358	15202.806	0.0134	0.246	-0.0316	2.304
Random Forest	-0.1527	17587.105	-0.106	0.276	-0.0985	2.539

Table 2: Performance of classifiers for number of upvotes

References

- Aashita Kesarwani. New York Times Comments, https://www.kaggle.com/aashita/nyt-comments, 2018.
- Sakshi Gupta. Predicting Popularity of The New York Times Comments. Towards Data Science, 2018.
- Yee Teh et al. Hierarchical Dirichlet Process. Berkeley, 2005.
- Bo Pang and Lilian Lee Thumbs up?Sentiment Classification using Machine Learning Techniques Proceedings of the Conference on Empirical Methods in Natural Language Processing
- Gensim: Topic modeling for humans Hierarchical Dirichlet Process. https://radimrehurek.com/gensim/models/hdpmodel.html