Structure, Actor or Culture: A Predictive Model on Online News Popularity

Weiwei Zheng
Master in Computational Social Science
University of Chicago
5801 S Ellis Ave, Chicago, IL 60637
Email: weiweiz@uchicago.edu

INTRODUCTION

According to the agenda setting theory, news influences what people think. In the time only with traditional media, people were passive recipients whose mindsets were highly manipulated by ideologies embedded in news articles. But thanks to the emerging new media nowadays, people attain more autonomy to process and reproduce information while news institutes are delivering information.

How can we know the public agenda? Besides news content, participance of opinion leaders and audience's discussion structure might be important indicators. This study tries to build a predictive model on news article popularity using data collected from social media site **reddit.com** where users can post, upvote and comment on news articles written by professional journalists. The result shows **users' discussion network** is the most important predictor on news popularity compared with participance of opinion leaders and other textual information .

DATA

Raw data (n = 19869):

- **Kaggle** — r/worldnews title 2014 - 2015
- **Reddit** — posts, comments and users (API praw)
- **Newspaper websites** — text (API newspaper3k)

Preprocessing (n = 4584)

- **Text feature**
  1) **topic modeling** distribution (API genism)
  2) **sentiment** of title and body (API vadersentiment)
  3) other basic text features
- **Network structure**
  1) the entire comment network (directed)
  2) the largest strongly connected component
- **Important actors**
  link and comment karma of important nodes (API networkx)

Github: https://github.com/ZhengErWei/MACS30200proj

MODEL PERFORMANCE

	Best Model (alpha = 3)	Dummy Regressor (strategy = 'mean')
R square (testing set)	0.62	< 0
MSE	21212725.87	55271237.27

VARIABLES AND METHOD

Dependent variable — upvote of each post

Independent variables (n = 50)

Culture

***Basic text features* —** word and sentence count, word and sentence length, part of speech, and lexical diversity

***Topic modeling distribution* —** ten topics

***Sentiment analysis* —** text and title sentiment

Structure — nodes and edges, strongly/weakly connected network, density, transitivity, average shortest path, eccentricity

Actor — actors with high betweenness, closeness, in/out degree centrality, and high and authority scores

Method — Ridge regression

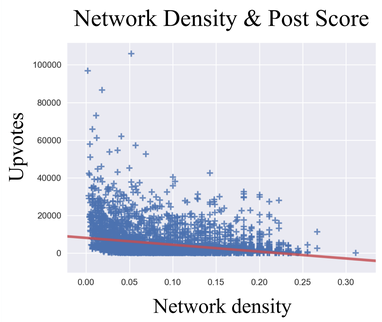
$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| \;=\mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|$$

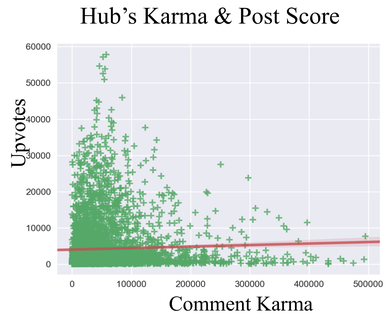
— Lasso regression

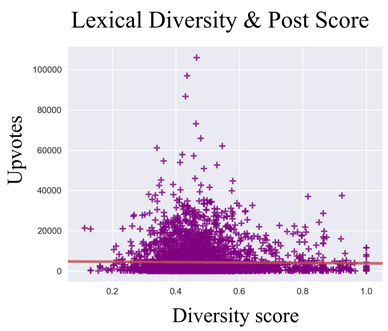
$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| \;=\mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|$$

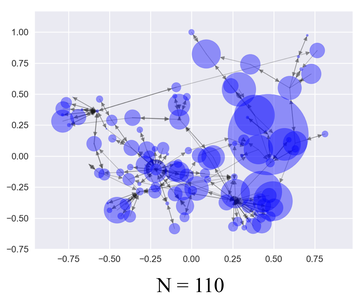
RESULTS

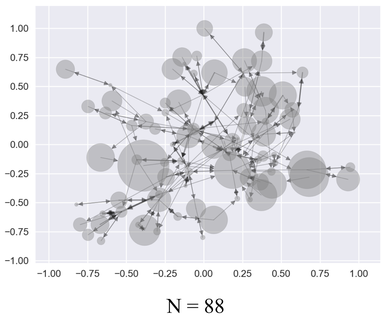
R square (testing set)	Structure model (n = 12)	Text model (n = 20)	Opinion leader model (n = 18)	Whole model (n = 50)
Ridge	0.611	-0.004	0.021	0.614
Lasso	0.611	-0.005	0.022	0.616

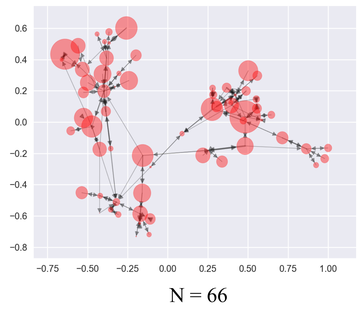












NON-ZERO PREDICTORS (n = 35)

Variable (best model)	Coefficient
# strongly connected components	55859.230
# nodes (whole network)	48565.011
# nodes (main component)	-28422.257
network diameter	-5677.734
variance of network eccentricity	4877.000
network density (main component)	-3181.042
network average shortest path	-2701.919
network hubs' comment karma	2485.400
author's link karma	1977.233
network center's comment karma	1842.214
author's comment karma	1527.211
average word length	1486.610
# weakly connected components	1193.473
percentage of adjectives	1170.682
topic: Malaysian Airline	1145.945
network density (whole network)	-1060.511
# sentences	-1010.400
mean of eccentricity	818.705
nodes of high closeness centrality	818.096
lexical diversity	-755.354
title sentiment	-751.607
topic: Russia	-519.945
topic: natural science	483.264
percentage of verbs	423.207
network authorities' comment karma	-418.811
topic: energy and mineral resources	376.802
topic: health and medicine	-245.991
sentence length	216.691
topic: nuclear and middle east war	-178.578
topic: cyber security	141.957
network transitivity	-120.060
text sentiment	63.635
title length	43.186
network authorities' link karma	-30.603

CONCLUSION

- **Network structure** is the most important predictor
- Participance of opinion leaders also matters
- Text features relatively trivial — challenges the results of most recent studies on online news popularity

NEXT STEP

- Interactive effect of the three dimensions
- Generality of the model on other Internet studies