Upvoted and Unbiased?

A textual analysis of “Top” vs “Controversial” Reddit headlines

IW03 - Computer Science Tools and Techniques for the Digital Humanities

Jessica Zheng, Advisor: Professor Brian Kernighan

Background on Reddit

- Reddit: A social news aggregation site.
  - divided into “subreddits” for various topics: politics, news, etc.
- User curation of content via up/down votes
  - More upvotes -> more visibility
- Content sorting options:
  - “Top”: highest measure of: (# upvotes - # downvotes)
  - “Controversial”: equal number of upvotes and downvotes

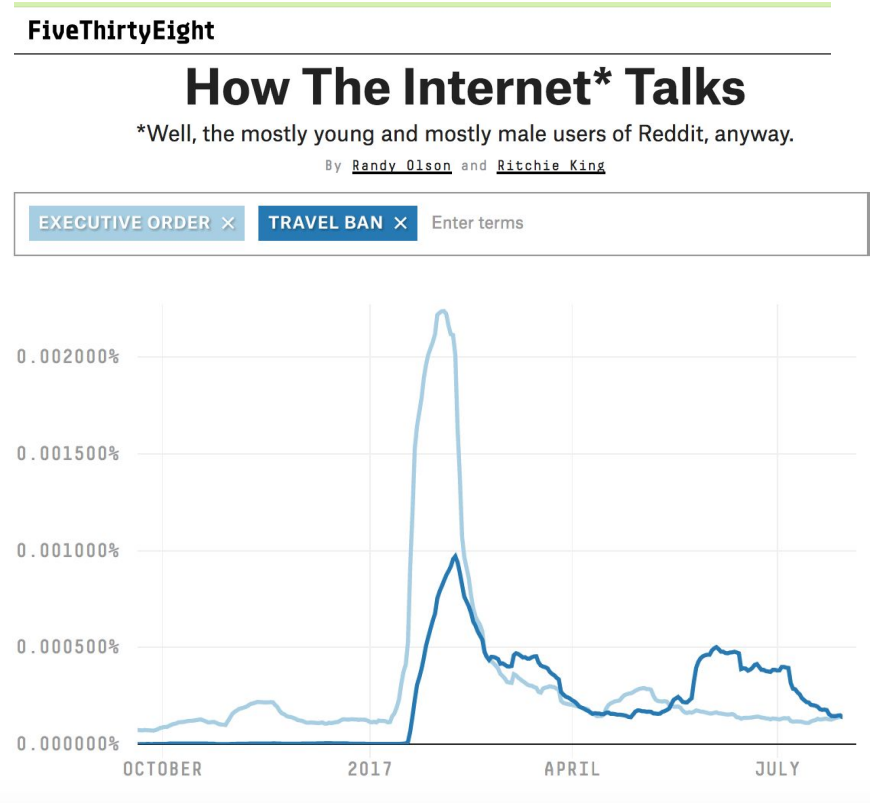
Motivation and Goal.

Motivation: How does **top** and **controversial** Reddit content tend to differ?

Goal: Create a data visualizer that compares “Top” and “Controversial” Reddit headlines over a 50 day period.

- how did Reddit treat major events over the data collection period?
  - Las Vegas Shooting, Net Neutrality Debate, Harvey Weinstein allegations...

Problem Background and Related Work



Existing visualizers track frequency of keywords, but don't factor in post popularity

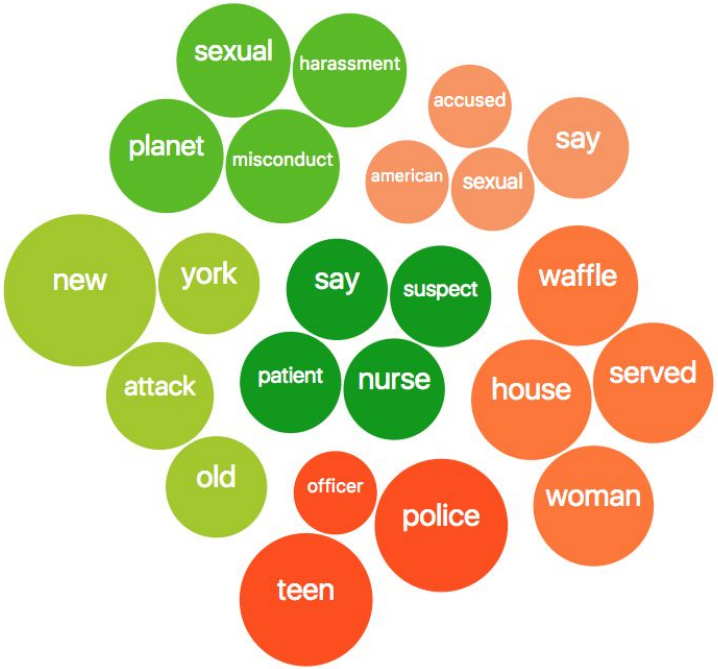
Approach

- Subreddits Used: r/politics, r/news, r/worldnews, r/technology
- Data collection period: October 1st 2017 - November 18th 2017
- Analyze **Top** and **Controversial** Reddit headlines via:
  - topic modeling (finding key topics present)
  - positive/negative sentiment
  - content source
- Observe patterns between analysis measures and indicators of content popularity (“top” or “controversial”, post score, # comments)

Dataset shown: r/news on 11/1/2017

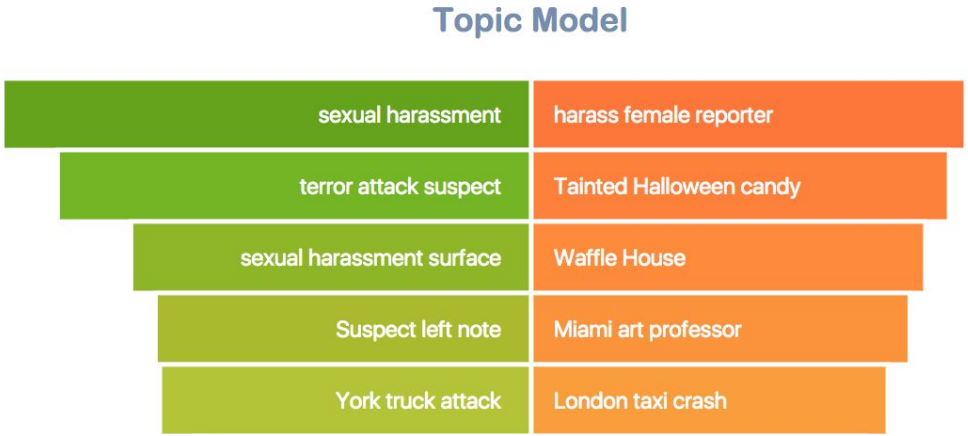
Topic Modeling

NLTK Latent Dirichlet Allocation (LDA)



^ randomized model: problematic!

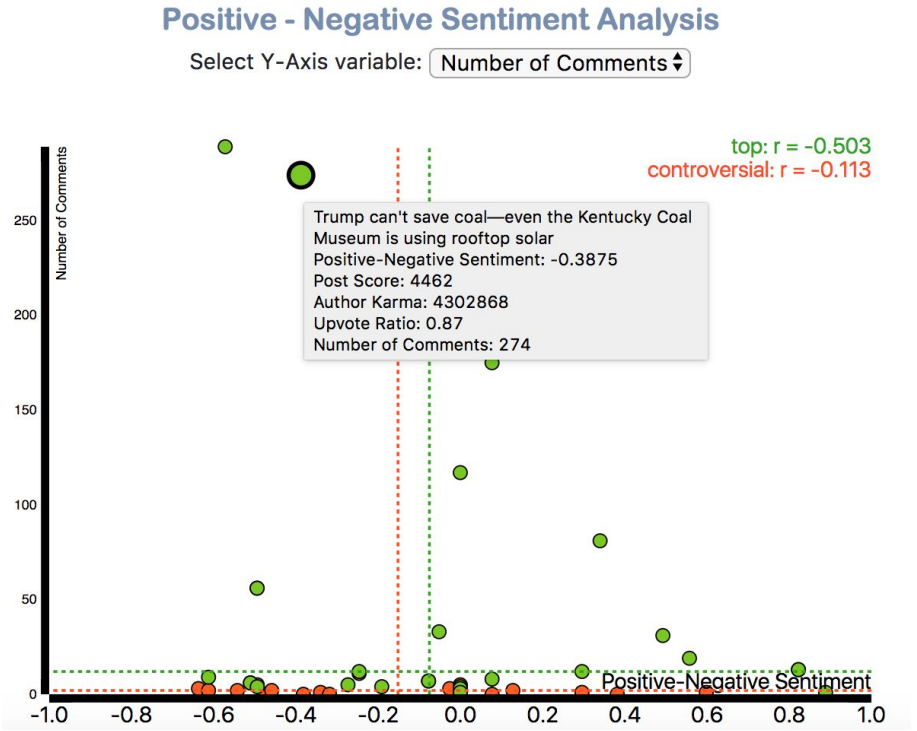
IBM Watson



returned more coherent topics

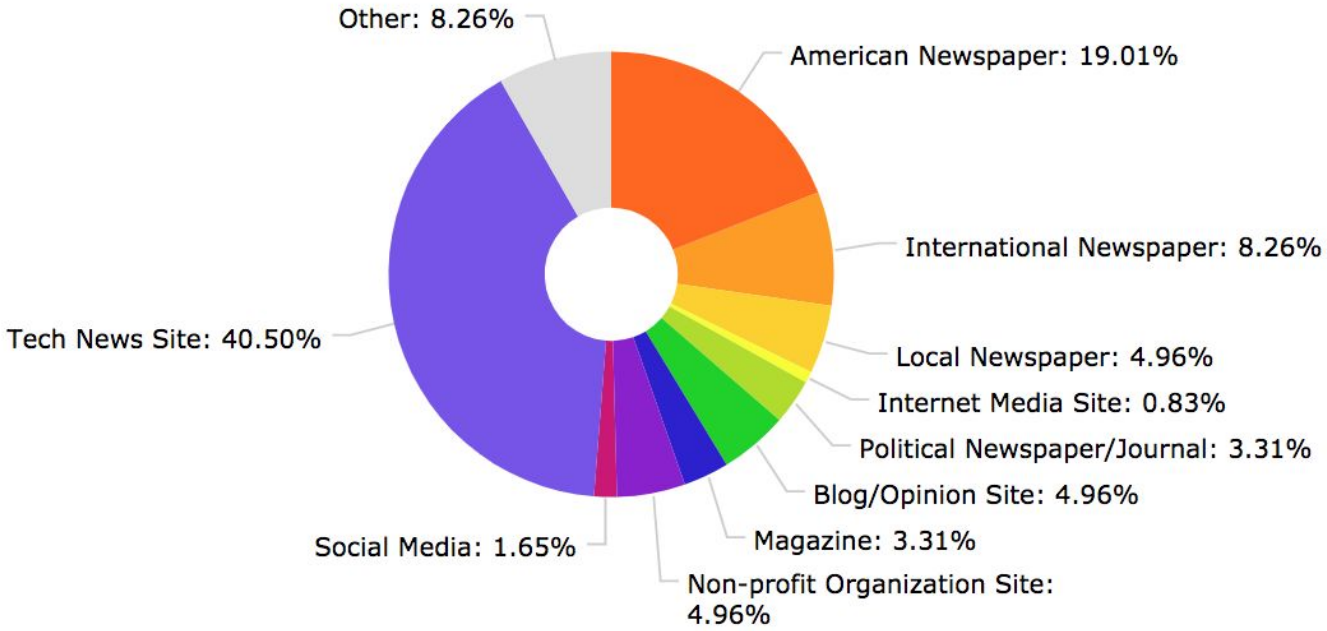
Positive/Negative Sentiment Analysis

- NLTK Vader
- Challenge: Vader didn't handle negative operators => some false neutral scores
  - "Patients *unwillingly* treated with electroshock at Norwegian mental hospitals."
- IBM Watson did take negative operators into account
- Solution: re-run all neutral-scored headlines through IBM Watson



Content Source Analysis

Parsed URLs, manually categorized most frequently appearing domains.



Results: Interesting Findings

r/politics Las Vegas Shooting

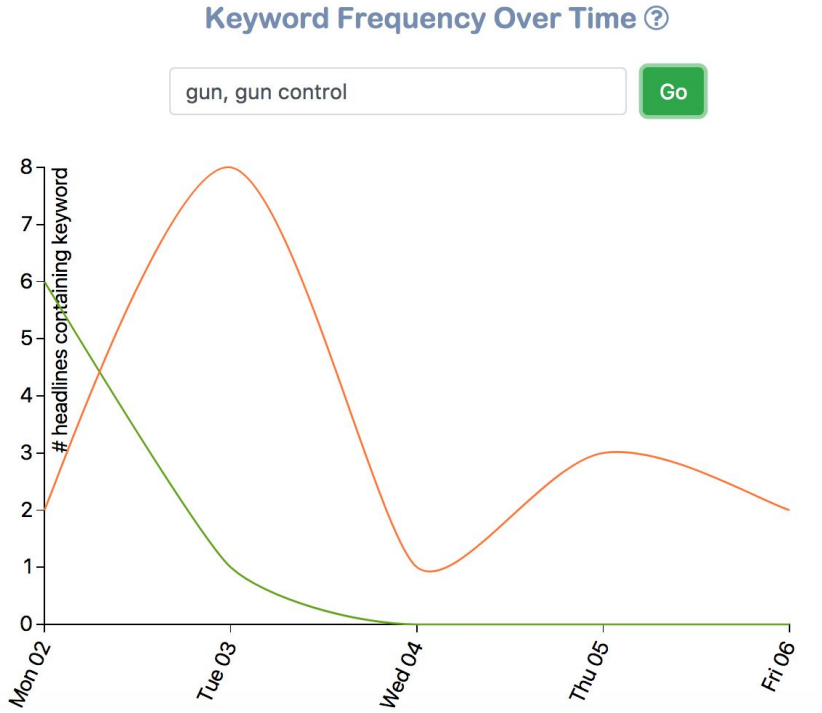
r/politics on 10/2/17-10/6/17

Showing data for 10/02/2017

Las Vegas shooting	Las Vegas
Puerto Rico	Las Vegas shooting
gun control	Trump
Las Vegas Gun	Trump tweets
Aid Shooting Victims	Trump calls
Las Vegas Official	beaten Trump
bigger domestic terrorist	pure evil

Showing data for 10/03/2017

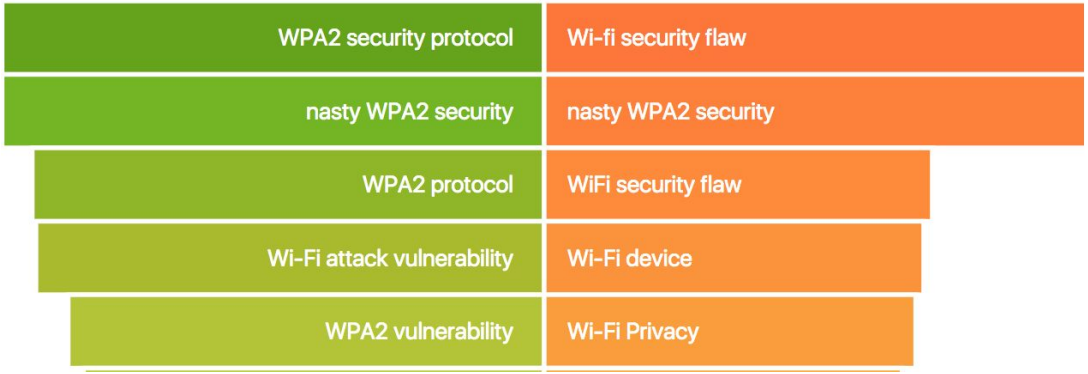
Puerto Rico	Las Vegas
Trump	'Republican gun toters
Pay Donald Trump	Trump tax cut
White House	Clinton talks gun
Ivanka Trump	gun control
Puerto Rico mayor	gun control talk
Trump aides	Las Vegas massacre



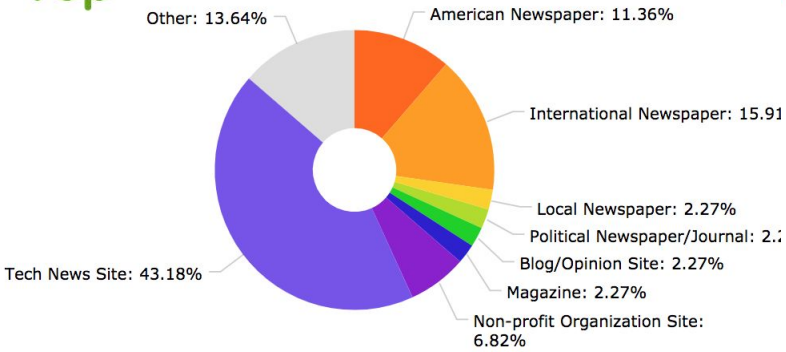
r/technology WPA2 vulnerability

Topic Model

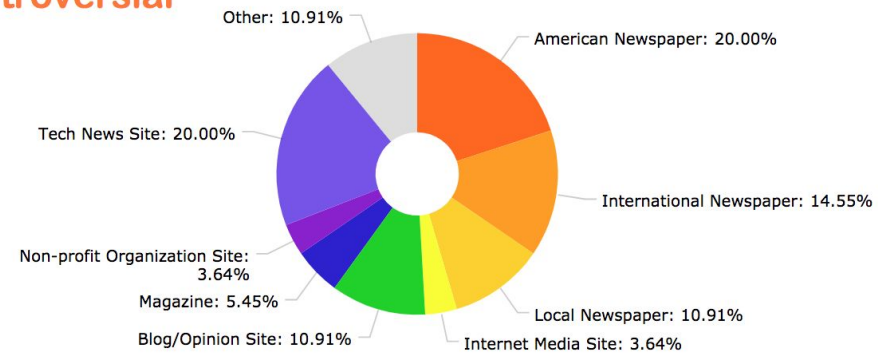
r/technology on 10/16/2017



Top



Controversial



Results - User Studies [In Progress]

Effectiveness of Analysis Tool:

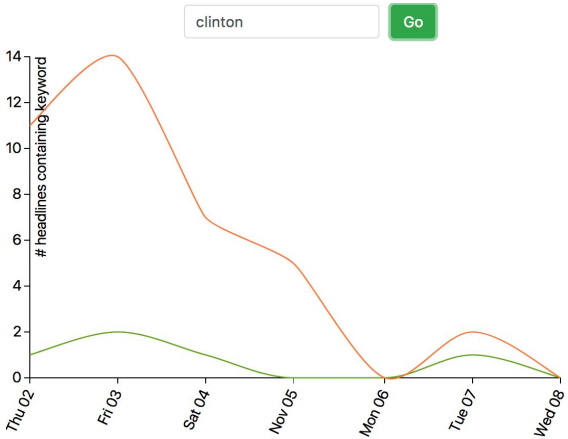
- 4 out of 5 users drew same conclusions I did for Las Vegas event
- 5 out of 5 users drew same conclusions I did for WPA2 event

Feedback on Usability:

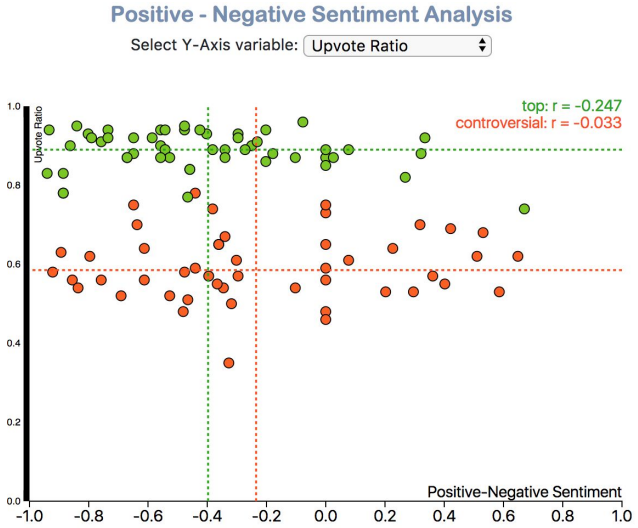
- requested: keyword frequency feature (like an ngram viewer)
- requested: hovercards explaining each analysis section
  - how the displays were computed
  - for scatterplots: what positive and negative x values mean

Other notable trends...

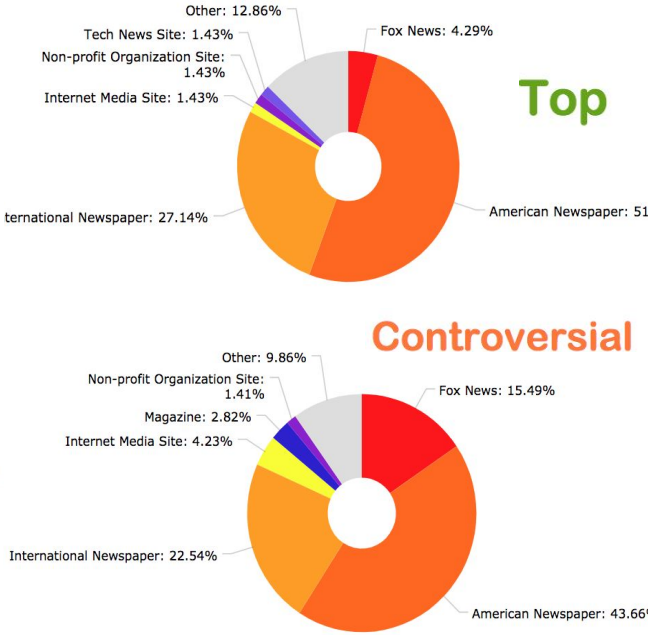
r/politics
controversial talks much more about clinton



r/news
top is on average more negative than controversial



r/news
controversial has more sources from Fox News



Conclusions

- Topic model and keyword frequency was most informative
- Positive-Negative sentiment was less insightful
- Implications:
  - For known controversial topics, Reddit moderators can intervene and promote healthy discussion.

Acknowledgements

- Professor Brian Kernighan - Independent Work Advisor
- Meagan Wilson - Independent Work Seminar TA
- Julie Zhu - Collaborator on Reddit-focused projects
- Dr. Nathan Matias - CITP researcher and Reddit Expert
- Professor Christiane Fellbaum - Natural Language Processing Expert
- Sources of code and data:
  - Data Collection - Python Reddit API Wrapper (PRAW)
  - Analysis Tools - NLTK, IBM Watson
  - Data Visualizer - Flask, PostgreSQL, Heroku, D3.js