Upvoted and Unbiased?

A textual analysis of "top" vs "controversial" Reddit headlines

Jessica Zheng '19 Advisor: Professor Brian Kernighan

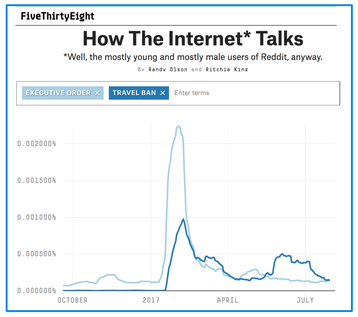
PRINCETON UNIVERSITY

1. Motivation

Goal: Create a data visualizer that compares and contrasts **Top-rated** and **Controversial-rated** Reddit headlines.

- Users draw their own conclusions about how Reddit treated major events over the data collection period.
  Las Vegas shooting, Net Neutrality debate etc.
- Existing Reddit visualizers do not account for post popularity, instead consider all posts.

  Problem: Statistics are computed from content that is not meaningful or impactful on Reddit users – as few as 5% of submitted content actually becomes popular.



2. Approach

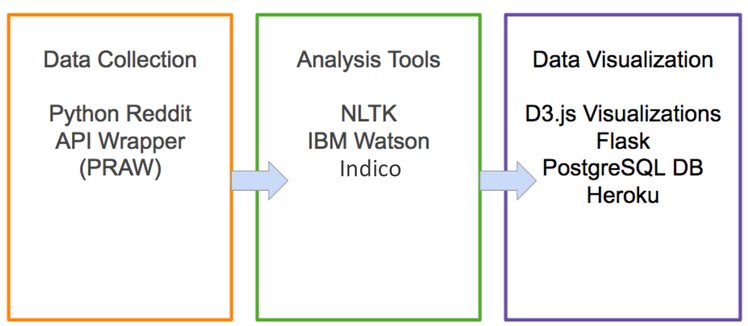
Analyze **Top-rated** and **Controversial-rated** Reddit headlines via:

- topic modeling
- positive-negative sentiment analysis
- political sentiment analysis
- content source breakdown
- keyword frequency

Subreddits used: r/politics, r/news, r/worldnews, r/technology
Data collection period: October 1st 2017 to November 18th 2017.

Observe patterns between analysis measures and indicators of content popularity ("top" or "controversial", post score, number of comments)

3. Implementation

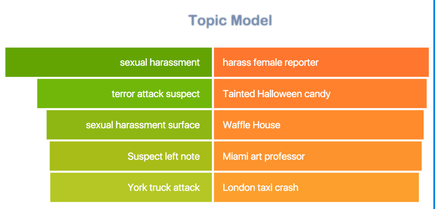


4a. Topic Modeling

NLTK Latent Dirichlet Allocation (LDA)

^ randomized model: problematic!

IBM Watson



returned more coherent topics

4b. Positive-Negative Sentiment Analysis

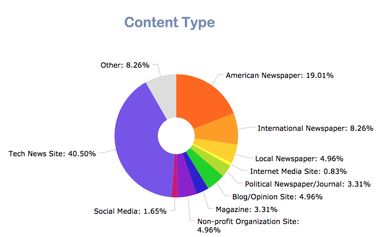
- NLTK's VADER Sentiment Analysis was a good starting point
- **Challenge:** VADER does not handle negative operators
  E.g. "Patients *unwillingly* treated with electroshock at Norwegian mental hospitals" predicted a neutral (zero) score
- **Solution:** Re-run neutral-scored headlines through IBM Watson's Sentiment Analyzer (which handled negative operators) to obtain a second measurement.

4c. Political Sentiment Analysis and Shortcomings

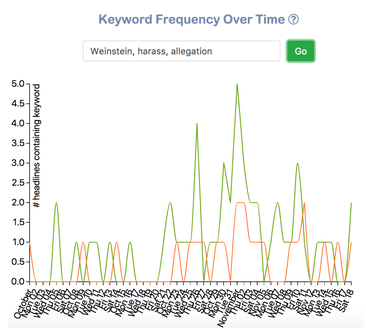
- Indico.io Text Analysis Service's API had limited accuracy
  - Negative operators: "I *don't* care for Trump." scores more highly on conservative leaning than liberal leaning.
- Somewhat better at detecting liberal/conservative people and topics, rather than liberal-conservative leanings
- Machine learning has lots of room for improvement!

4d. Other Analysis Measures

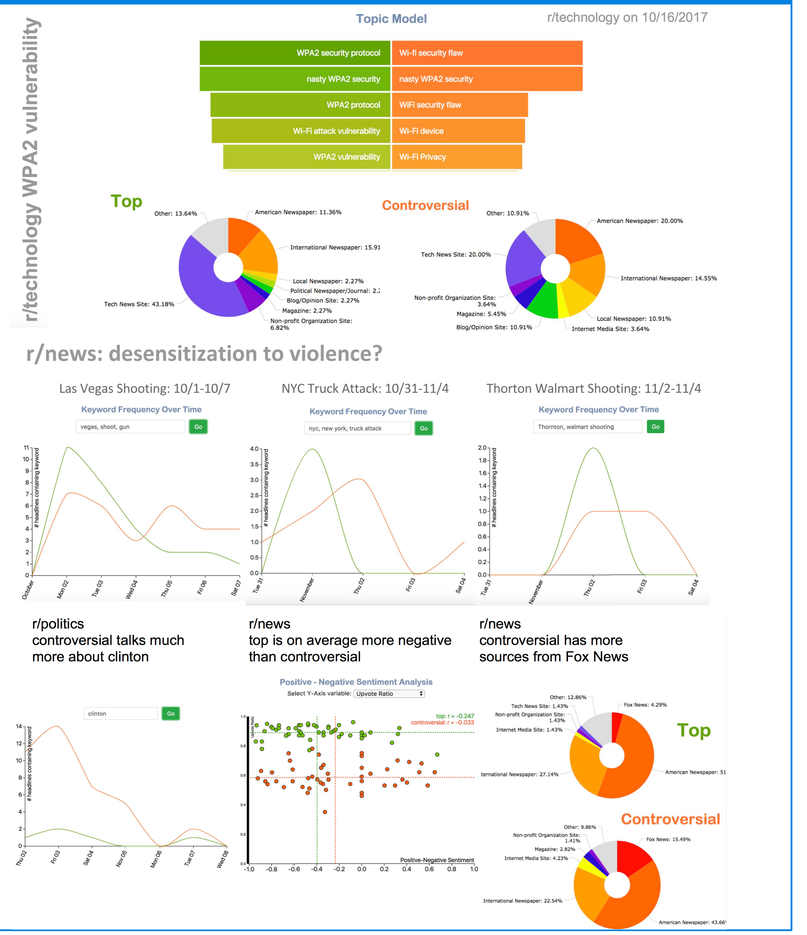
Content Source Breakdown



Keyword Frequency in Headlines



5. Findings



6. Usability Evaluation

- Effectiveness of analysis tool:
  - 4/5 users drew the same conclusions I did for Las Vegas Event
  - 5/5 users drew the same conclusions I did for the WPA2 event
- Usability issues were mostly concerned with learnability (easily resolved) and limitations of the sentiment analysis tools.

7. Conclusions and Future Work

- Topic model and keyword frequency visuals were most informative
- Machine learning still has room to go for sentiment analysis
- Implications:
  - For known controversial topics, Reddit moderators can intervene and promote healthy discussion
- Future Work: Combine multiple analysis forms in the same visual, create improved sentiment classifiers, explore more subreddits.