1. Motivation

Goal: Create a data visualizer that compares and
Top-rated and **Controversial-rated** Reddit he

- Users draw their own conclusions about how

Upvoted ar

A textual analysis of "top" vs '

Jessica Zheng '19 **Advi**

contrasts

adlines.

Reddit treated

4a. Topic Modeling

NLTK Latent Dirichlet Allocation (LDA)

ıd Unbiased?

"controversial" Reddit headli

isor: Professor Brian Kernighan

IBM Watson

Topic Model

sexual harassment | harass female reporter

5. Findings

2 vulnerability

nes

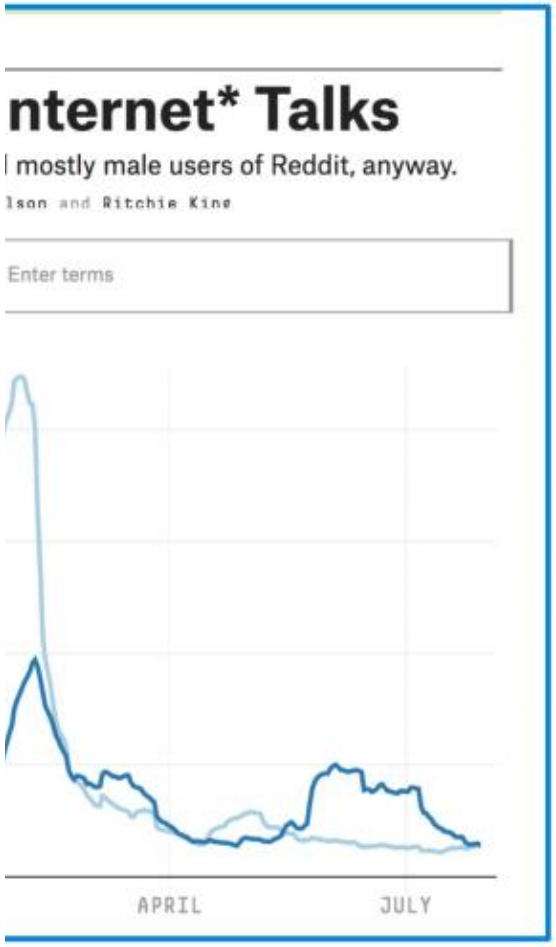

Topic Model
r/technology on 10/16/2017
WPA2 security protocol
Wi-fi security flaw
nasty WPA2 security
nasty WPA2 security
WPA2 protocol
WiFi security flaw
Wi-Fi attack vulnerability
Wi-Fi device

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



^ randomized model: problematic!

4b. Positive-Negative Sentim

- NLTK's VADER Sentiment Analysis
- **Challenge:** VADER does not handl
  E.g. "Patients *unwillingly* treate
  Norwegian mental hospitals" p
- **Solution:** Re-run neutral-scored h

terror attack suspect	Tainted Halloween candy
sexual harassment surface	Waffle House
Suspect left note	Miami art professor
York truck attack	London taxi crash

returned more coherent topics

nent Analysis

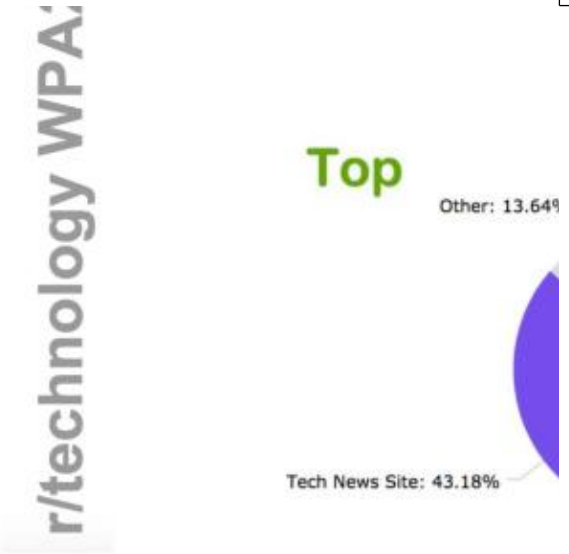
was a good starting point

e negative operators

ed with electroshock at

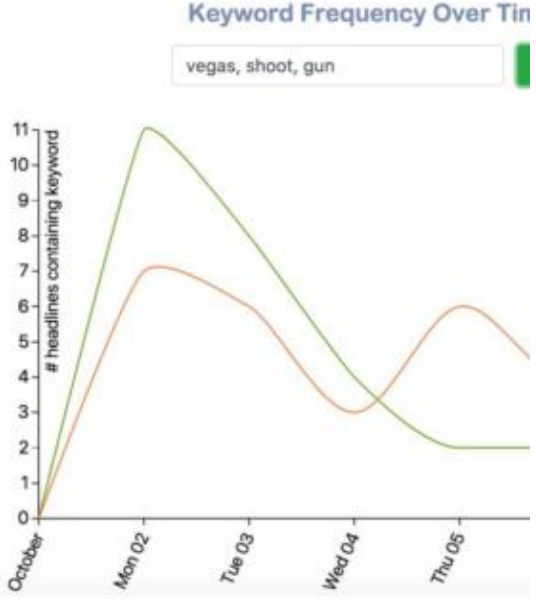
redicted a neutral (zero) score

eadlines through IBM Watson's



r/news: desensiti

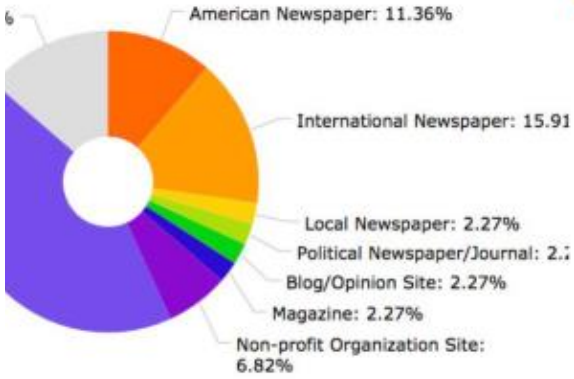
Las Vegas Shooting: 10/

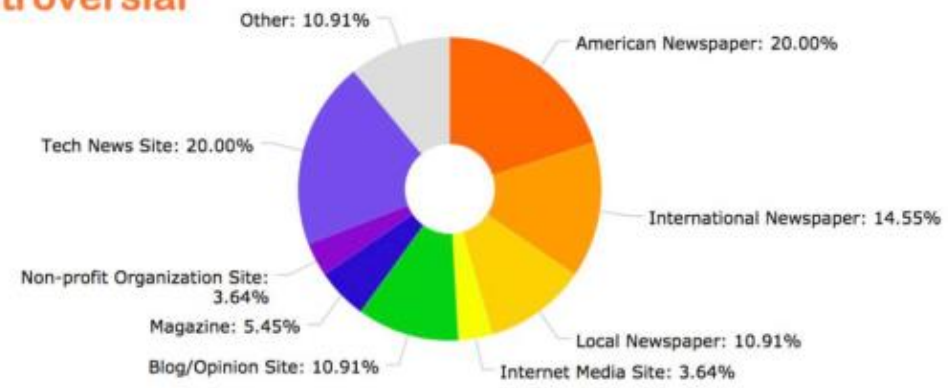


r/politics

WPA2 vulnerability | Wi-Fi Privacy

Controversial

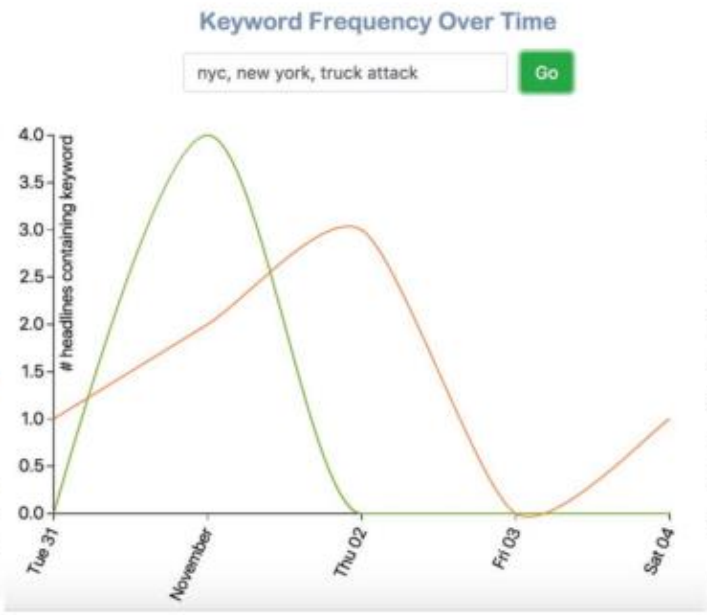




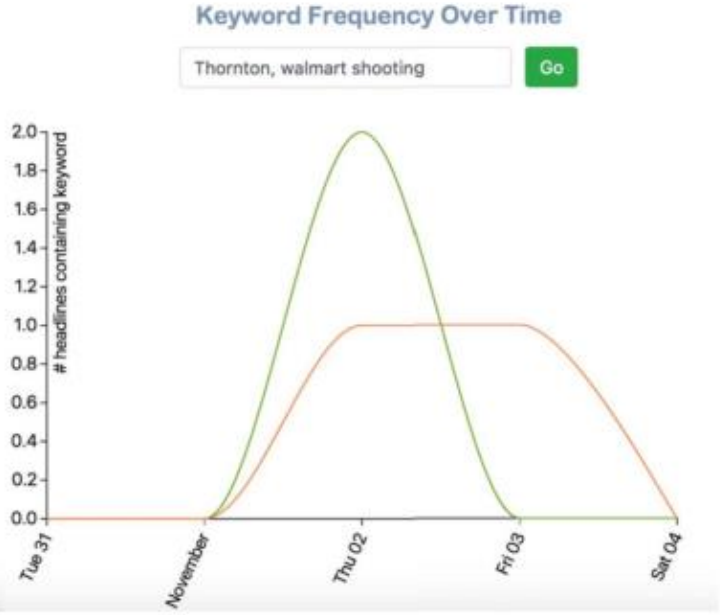
ization to violence?

1-10/7

NYC Truck Attack: 10/31-11/4



Thorton Walmart Shooting: 11/2-11/4



r/news

r/news

Analyze **Top-rated** and **Controversial-rated** Reddi

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

Subreddits used: r/politics, r/news, r/worldnews, r

Data collection period: October 1st 2017 to Novem

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

3. Implementation

t headlines via:

/technology
ıber 18th 2017.

indicators of
score, number

Sentiment Analyzer (which handle
a second measurement.

4c. Political Sentiment Analy

- Indico.io Text Analysis Service's AI
  - Negative operators: "I *don't* ca
    highly on conservative leaning
- Somewhat better at detecting libe
  topics, rather than liberal-conserv
- Machine learning has lots of roon

4d. Other Analysis Measure

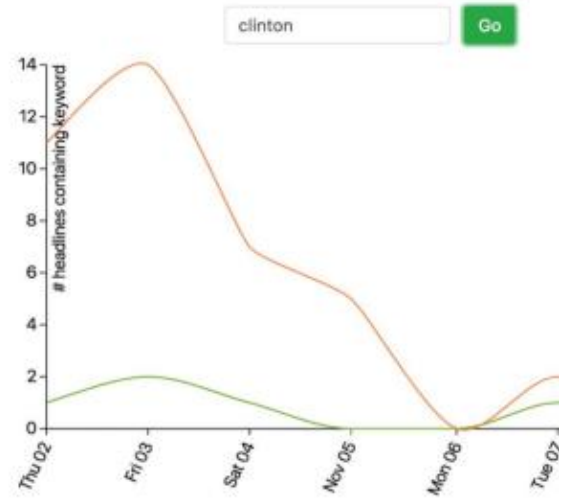
ed negative operators) to obtain

ysis and Shortcomings

PI had limited accuracy
re for Trump." scores more
than liberal leaning.
eral/conservative people and
vative leanings
n for improvement!

es

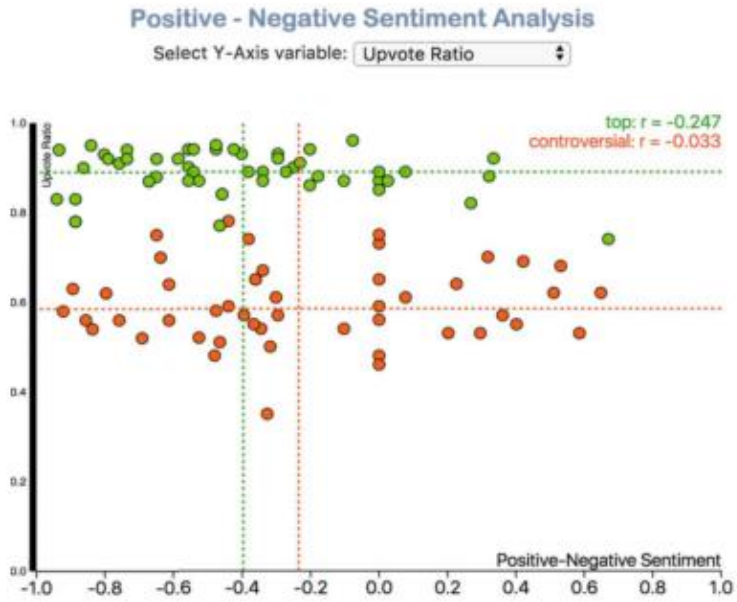
controversial talks much more about clinton



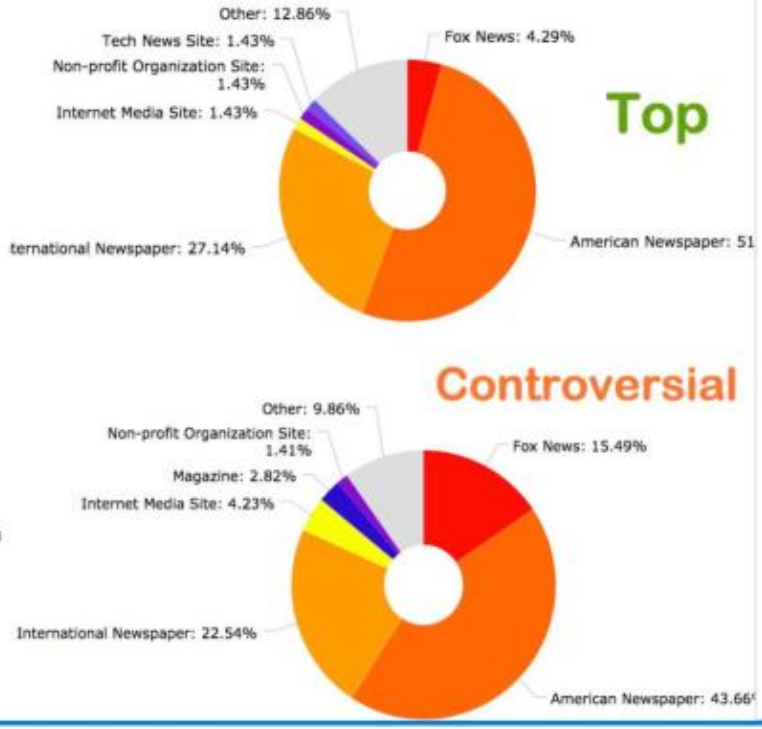
6. Usability Eva

- Effectiveness of a
  - 4/5 users drev
  - 5/5 users drev

top is on average more negative than controversial



controversial has more sources from Fox News

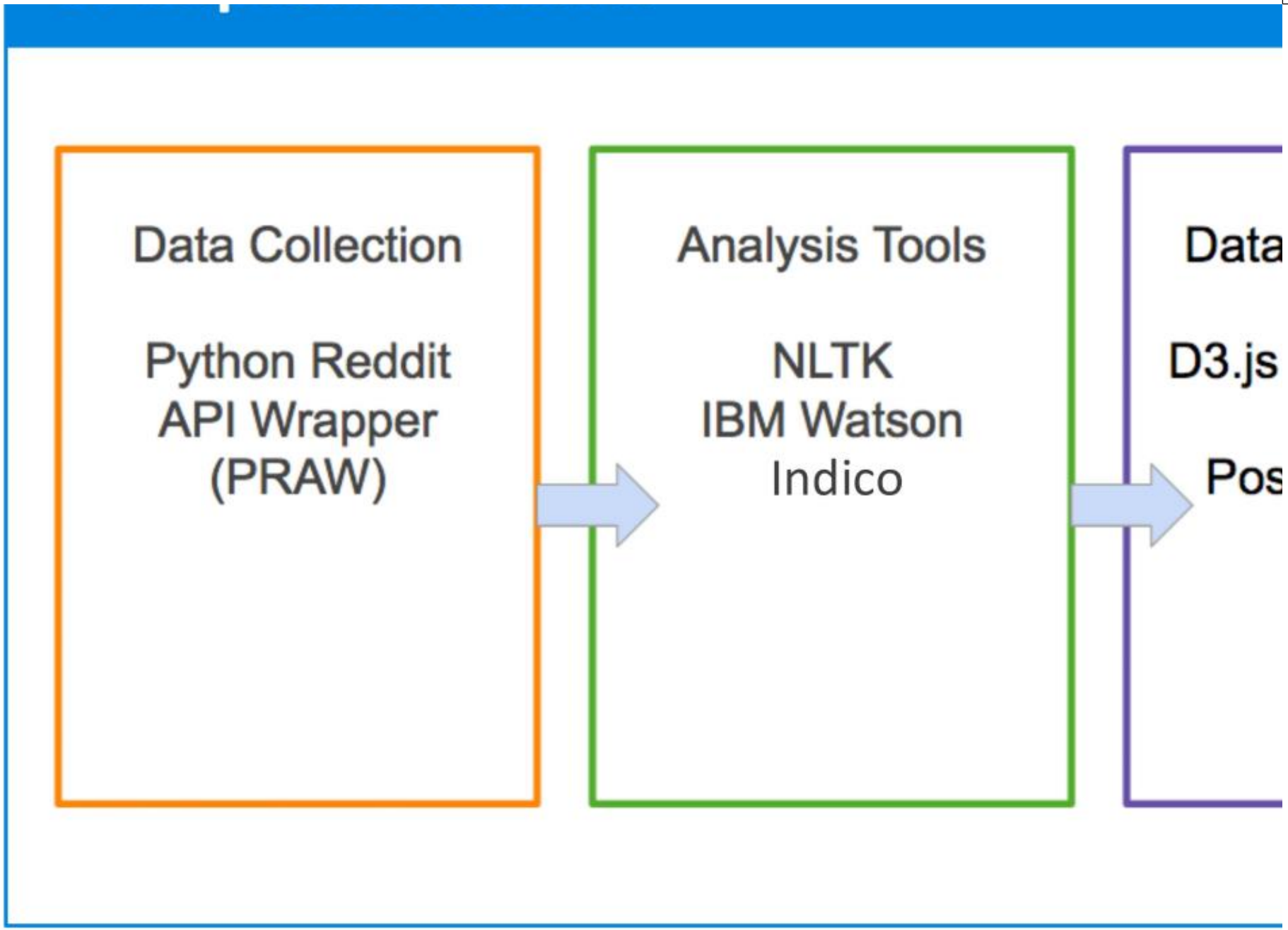


luation

nalysis tool:

v the same conclusions I did for Las Vegas Event

v the same conclusions I did for the WPA2 event

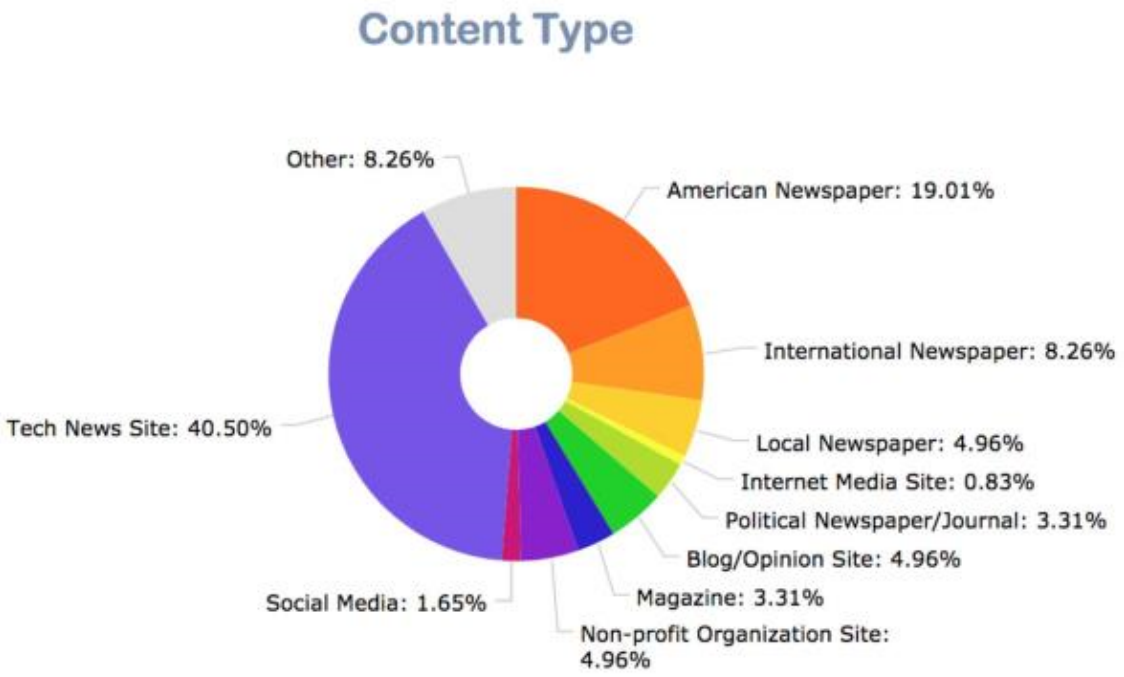

Data Collection
Python Reddit API Wrapper (PRAW)
Analysis Tools
NLTK
IBM Watson
Indico
Data
D3.js
Pos

Visualization

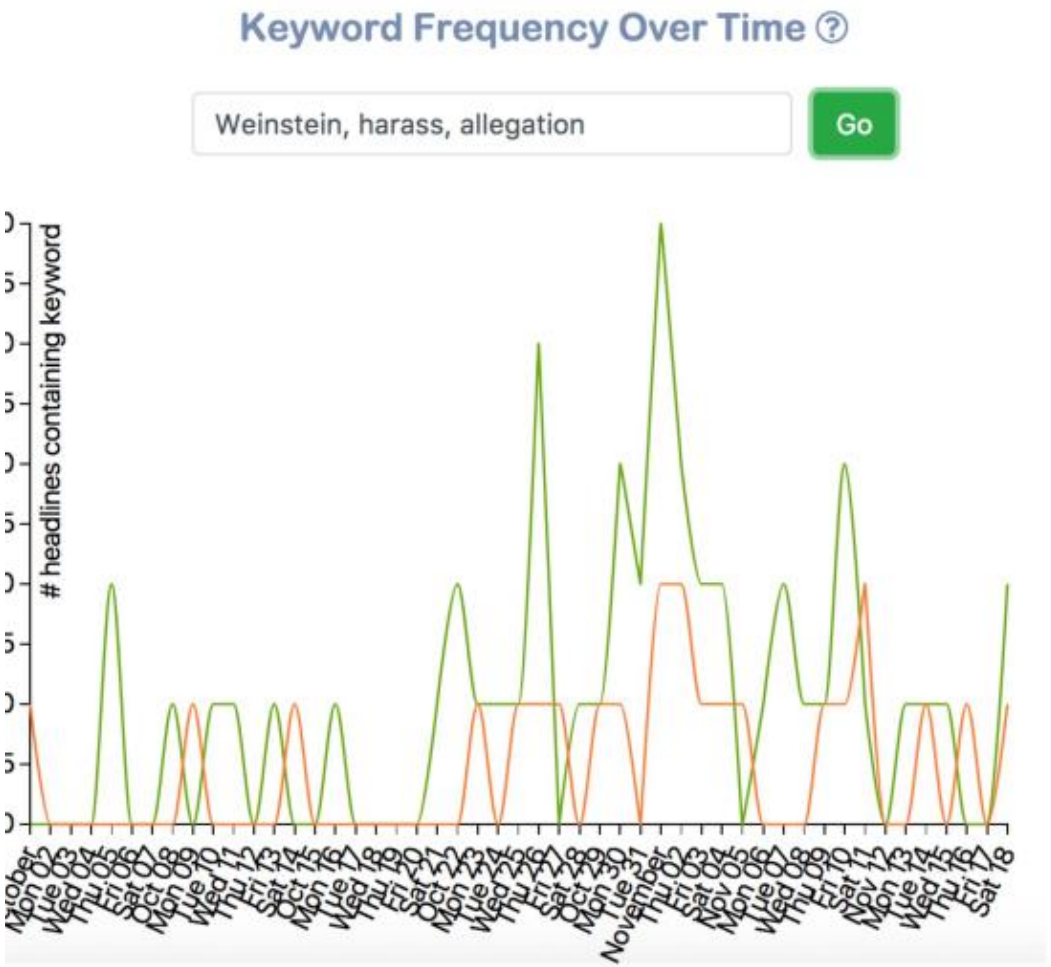
Visualizations
Flask
tgreSQL DB
Heroku

Content Source Breakdown

Ke



yword Frequency in Headlines



- Usability issues w
  resolved) and lim

7. Conclusions

- Topic model and
- Machine learning
- Implications:
  - For known cor
    and promote h
- Future Work: Cor
  create improved

ere mostly concerned with learnability (easily
itations of the sentiment analysis tools.

and Future Work

keyword frequency visuals were most informative
still has room to go for sentiment analysis

troversial topics, Reddit moderators can intervene
ealthy discussion
nbine multiple analysis forms in the same visual,
sentiment classifiers, explore more subreddits.