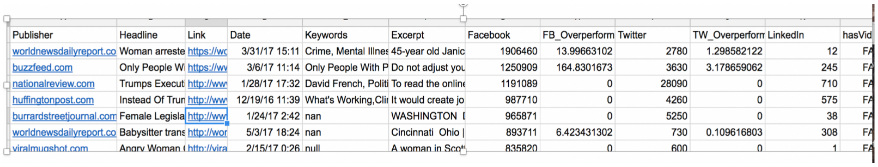
Wei Tang
Dec 26 · 3 min read

Sentiment Analysis

Headline analysis using TextBlob and NLTK:

An exploration 10,000 news articles collected between November 2016 and May 2017 from the top-500 news publishers.



Publisher	Headline	Link	Date	Keywords	Excerpt	Facebook	FB_Overperform	Twitter	TW_Overperform	LinkedIn	hasVid
worldnewsdailyreport.co	Woman arreste	https://wc	3/31/17 15:11	Crime, Mental Illnes	45-year old Janic	1906460	13.99663102	2780	1.298582122	12	FA
buzzfeed.com	Only People Wi	https://wv	3/6/17 11:14	Only People With P	Do not adjust you	1250909	164.8301673	3630	3.178659062	245	FA
nationalreview.com	Trumps Executi	http://ww	1/28/17 17:32	David French, Politi	To read the online	1191089	0	28090	0	710	FA
huffingtonpost.com	Instead Of Trun	http://ww	12/19/16 11:39	What's Working,Clir	It would create jo	987710	0	4260	0	575	FA
burrardstreetjournal.com	Female Legisla	http://ww	1/24/17 2:42	nan	WASHINGTON [	965871	0	5250	0	38	FA
worldnewsdailyreport.co	Babysitter trans	http://wor	5/3/17 18:24	nan	Cincinnati Ohio |	893711	6.423431302	730	0.109616803	308	FA
viralmugshot.com	Angry Woman (	http://vira	2/15/17 0:26	null	A woman in Scot	835820	0	600	0	1	FA

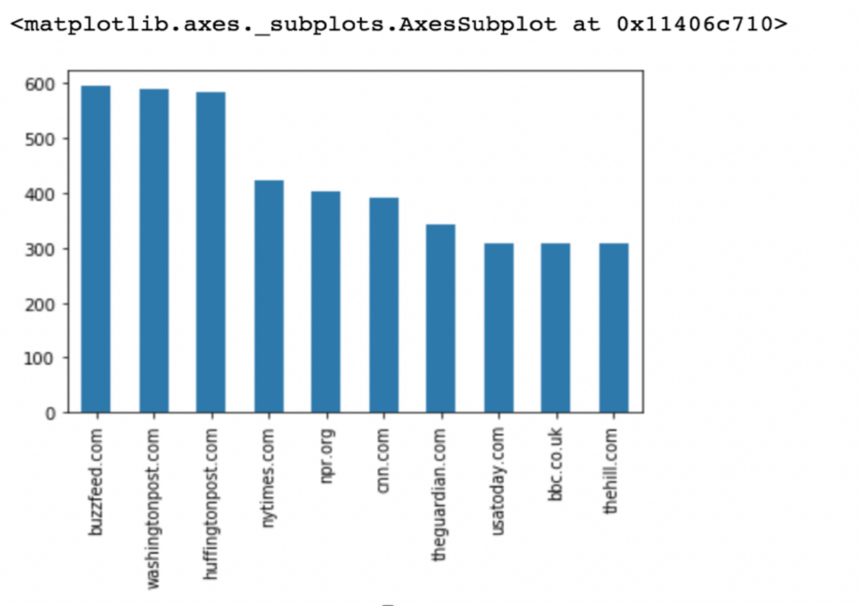
Here is how the dataset looks like

Data analysis always begins with questions.I was working with a dataset of 10,000 stories from the top-500 news publishers in the U.S. Before I could determine what to filter out for analysis, I had to define my questions.

Questions:

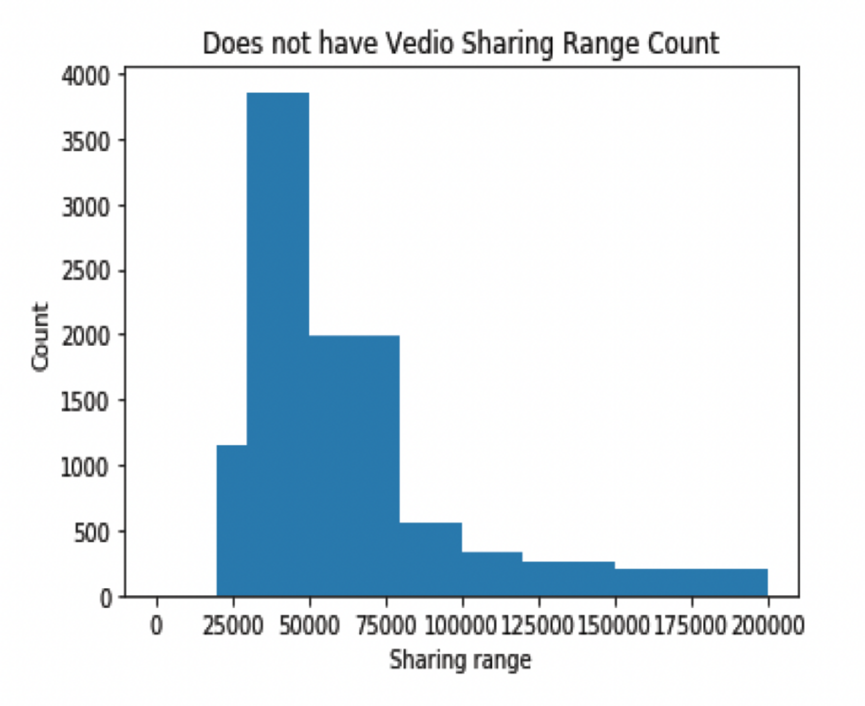
- Which news organizations publish the most articles ?
- What are the top keywords among all headlines.?
- Are stories with videos are more likely to do share? Or less?
- What is the most frequency words appear in 10k articles?
- What the proportion of positive news as well as the negative news.
- Can we find a specific linguistic pattern of the headline of every publisher?

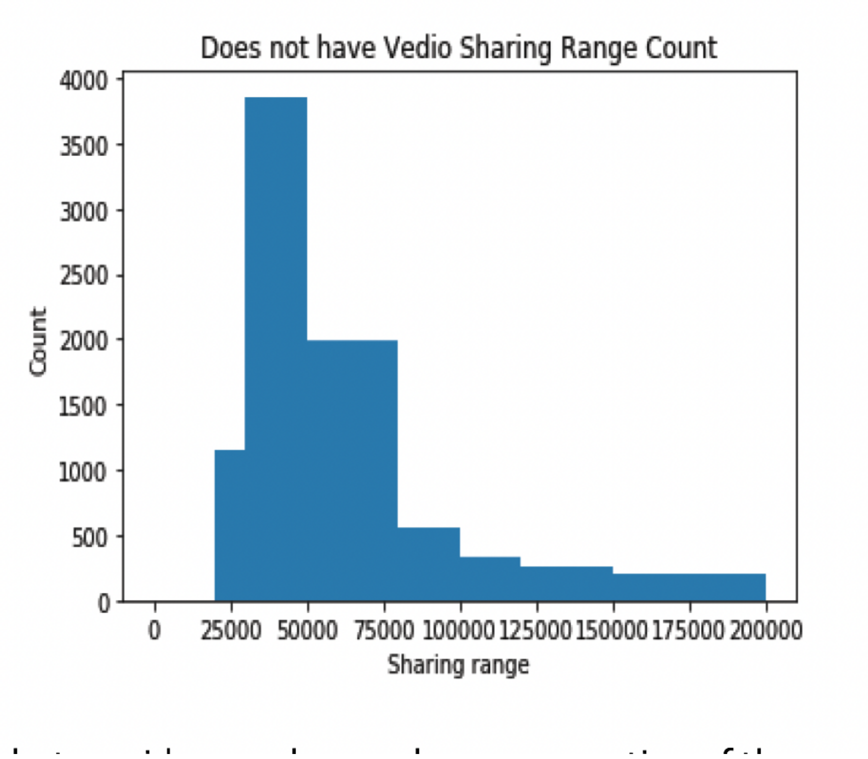
Which publisher published the most articles?



```
<matplotlib.axes._subplots.AxesSubplot at 0x11406c710>
```

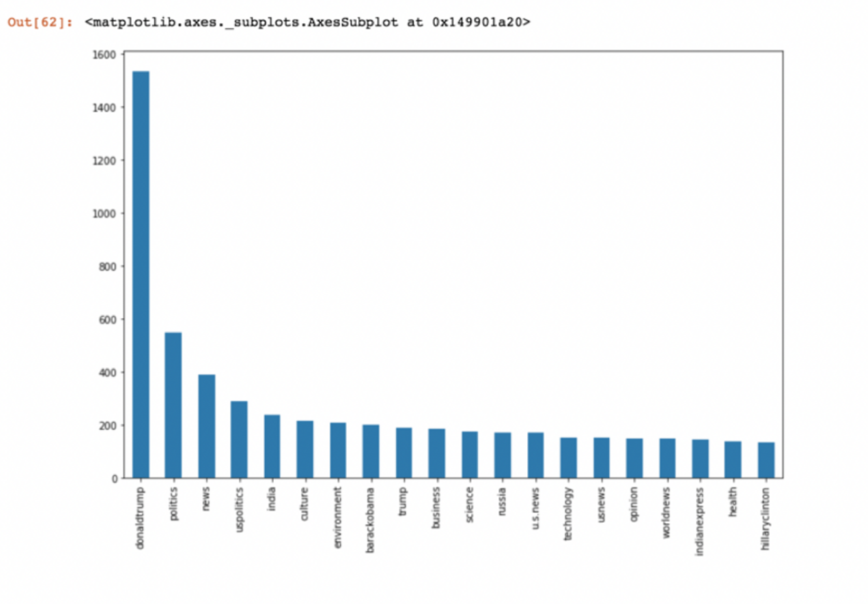
Are stories with videos more likely to do be shared?





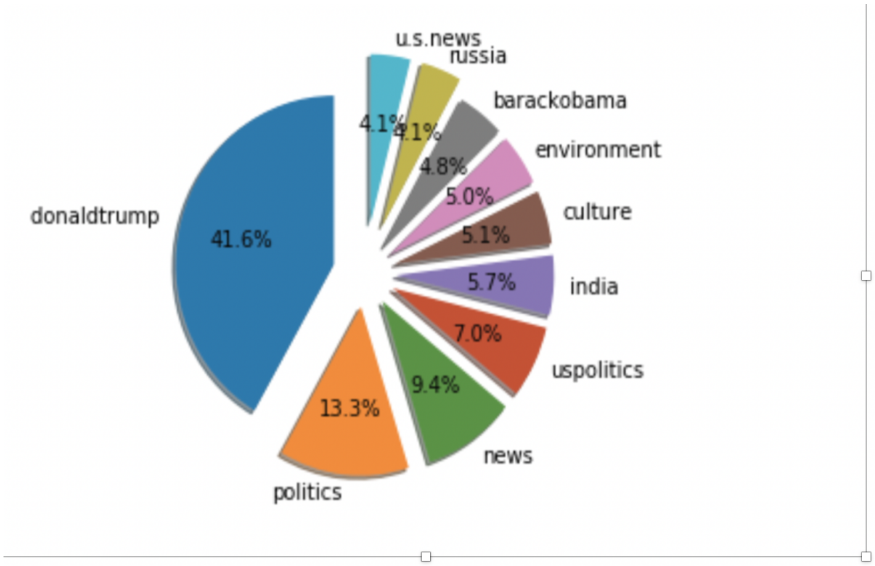
Conclusion:The sharing behavior appears to follow a similar pattern but no videos make up a larger proportion of the datasets.

Plotting the top 20 keywords

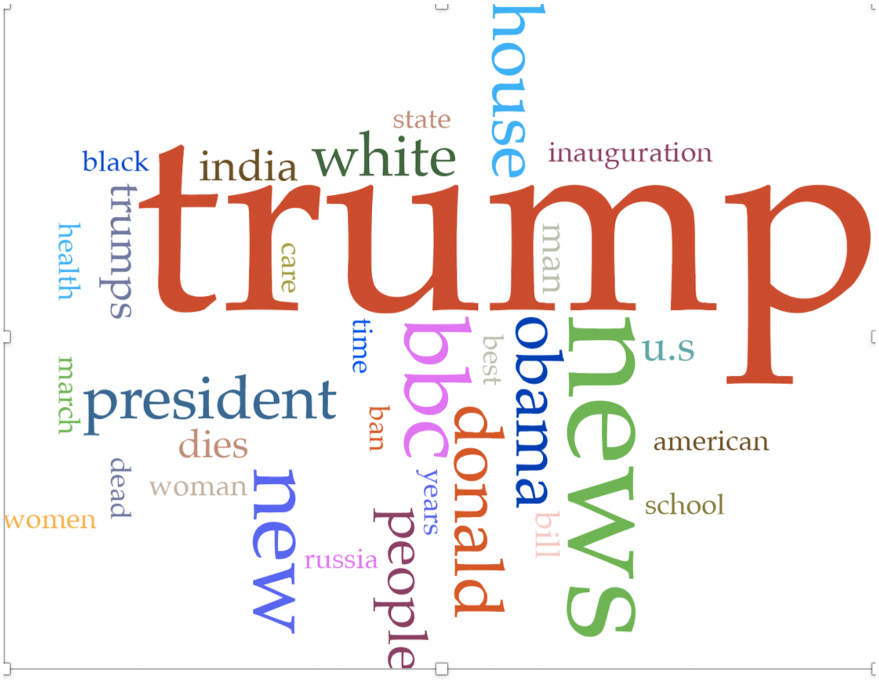


```
Out[62]: <matplotlib.axes._subplots.AxesSubplot at 0x149901a20>
```

Top 10 keywords from the column of Keyword



Word clustering 30 most frequent words appear in headlines



Most **frequent words** in the headline:

trump (1842);

news (588);

bbc (375)

donald (293);

house (281);

president (273);

obama(262);

white (255); people (244);

india (215);

u.s(198);

dies (196);

bill (194);

woman (157);

march (145);

russia (144);

health(131);

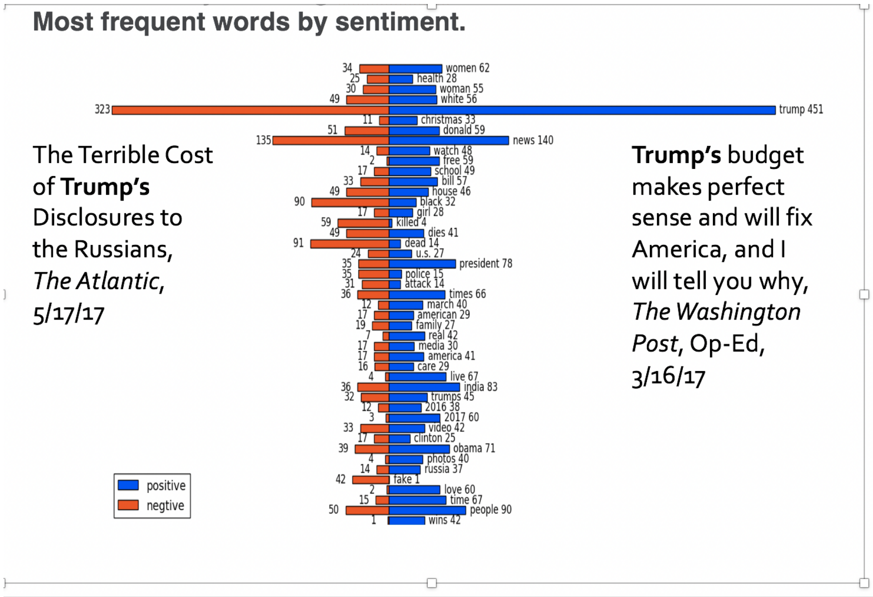
black (130);

inauguration (124);

school (108)

Sentiment Analysis

Headline analysis using TextBlob and NLTK:



Most frequent words by sentiment.

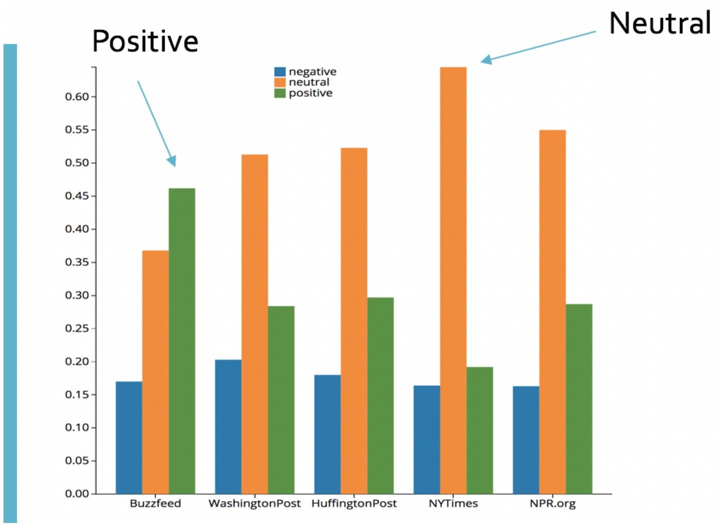
The Terrible Cost of **Trump's** Disclosures to the Russians, *The Atlantic*, 5/17/17

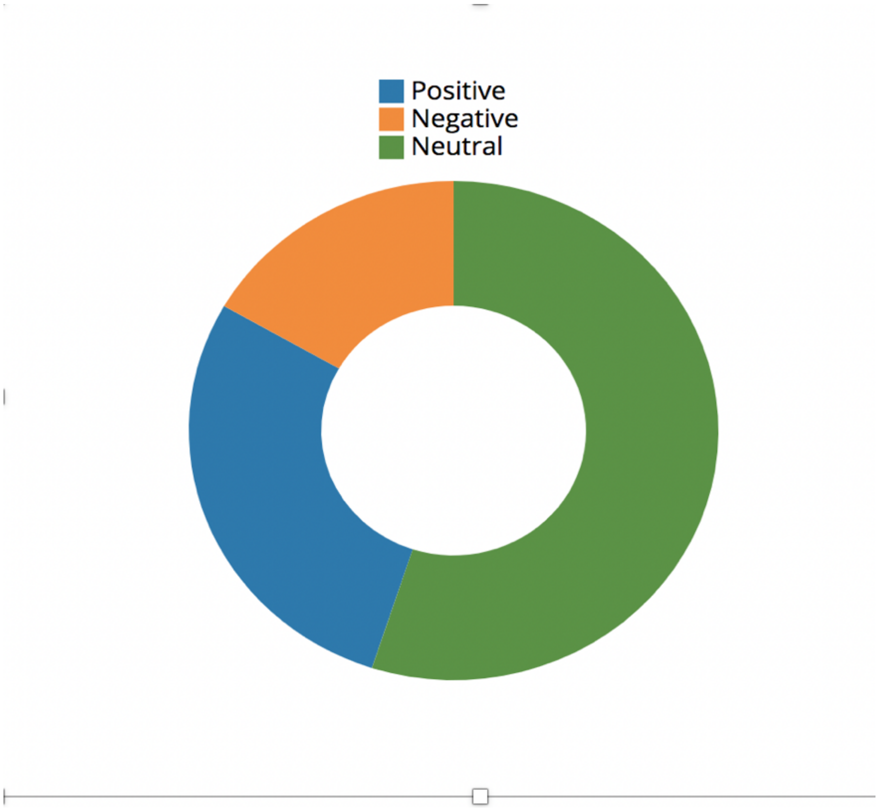
Trump's budget makes perfect sense and will fix America, and I will tell you why, *The Washington Post*, Op-Ed, 3/16/17

However, sentiment analysis is bad at catching sarcasm, for example:

> *Ruth Bader Ginsburg on Trumps presidency: We are not experiencing the best of times, The Washington Post*

> *Thank you, Trump voters, for this wonderful joke, The Washington Post*





overall headline sentiment analysis