Media Sentiment Report

Zehan Chao, Denali Molitor

March 31, 2020

1 Introduction

In this work, we aim to study the sentiment and ideology type from media documents. Currently we are working on Media Tweets classification and clustering based on their left/right ideology and source quality. We will describe several experiments done with our current dataset.

2 Experiments

2.1 Experiment Settings

We used MediaBiasChart V5.0 as ground-truth labeling. There are 90 media appeared in the MediaBiasChart at *https://www.adfontesmedia.com/interactive-media-bias-chart*.

We found the media tweets from another data source at *https://dataverse.harvard.edu/dataverse/gwu-libraries*. This dataset (NewsOutletTweet) contains 39,695,156 tweets collected from approximately 4,500 media twitter accounts. They were collected between August 4, 2016 and July 20, 2018.

For each media in MediaBiasChart, we manually look for its corresponding Twitter account and check if it's in the NewsOutletTweet. Finally, we obtained 65 medias appearing in both MediaBiasChart and NewsOutletTweet. The table on the right is a snap of the list of their twitter names. The full list could be found in file */code/MediaDataDescibe.ipynb*

The Bias/Quality score are from */data/MediaBiasChart.csv*. This csv file contains 10 to 20 reviewed articles with Bias/Quality score for each media, we grouped the review score according to their source media and took the average of each group served as Bias/Quality score of each media. The value is almost the same as the position of media on the MediaBiasChart

	Source	Bias	Quality
0	ABC	-1.846000	49.866500
1	AP	-1.063261	52.189130
2	axios	-5.737857	47.303571
3	BBC	-3.033333	46.266667
4	Bloomberg	-0.850345	47.522759
5	BreitbartNews	18.987857	30.637143
6	businessinsider	-0.378000	43.283333
7	BuzzFeed	-7.061333	43.167333
8	CBSNews	-1.846154	46.839231
9	CNN	-8.553827	40.487716
10	cnsnews	25.750000	27.750000
11	csmonitor	-0.205294	44.265294
12	thedailybeast	-12.039231	38.800769
13	DailyCaller	20.061333	28.800000
14	DailySignal	19.972667	30.410667
15	democracynow	-16.710667	37.544000
16	FT	0.621538	47.467692
17	TheFiscalTimes	1.522000	44.544667
18	Forbes	0.202857	39.845000
19	ForeignPolicy	-1.649333	41.692000

With the media twitter names, we filtered out 1417030 tweets which all come from the 65 media in the list. The frequency of tweets for each media is fairly balanced: except the Bloomberg (127 tweets), each media has between 5k and 70k tweets in this dataset.

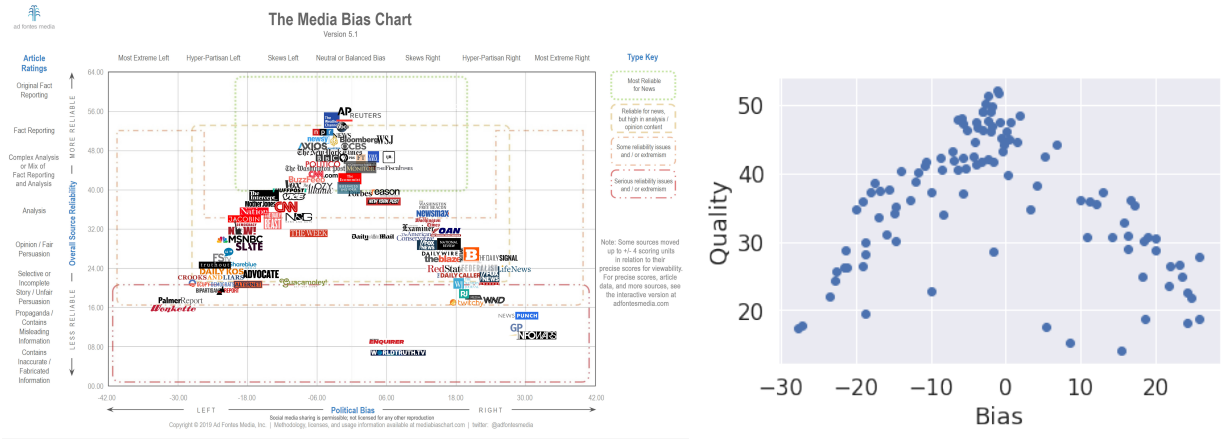
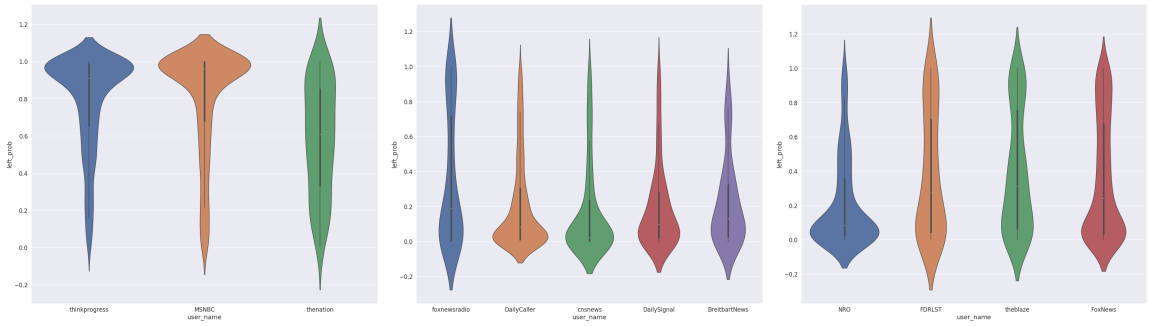


Figure 1: The original Media Bias Chart vs. their actual Bias/Quality

2.2 Binary Classification

The first experiment we tried is to reproduce a similar classification as Denali did for Democracy/Republic senator tweets. We labeled the media with political_bias ≤ -15 as left_media and political_bias ≥ 15 as right_media. Then randomly select half of the left_media, combine their tweets as train_left and obtained train_right in the similar way. With the naive Bayes model and the train dataset, we are able to predict the probability of each tweet being a 'left' tweet or 'right' tweet.

When we use all 1417030 tweets, the tweet distributions for each media did not really tell the media is left or right. i.e. nearly a half of the medias labeled as 'right' in MediaBiasChart are posting more 'left' tweets. Hence we conclude the binary model was not working well with the entire dataset. However, when we filtered the tweets with some keywords (for details, see */code/MediaRegression.ipynb*), The tweet distribution of each media is much clearer:



(a) train_left (b) train_right (c) test_right

Figure 2: The original Media Bias Chart vs. their actual Bias/Quality

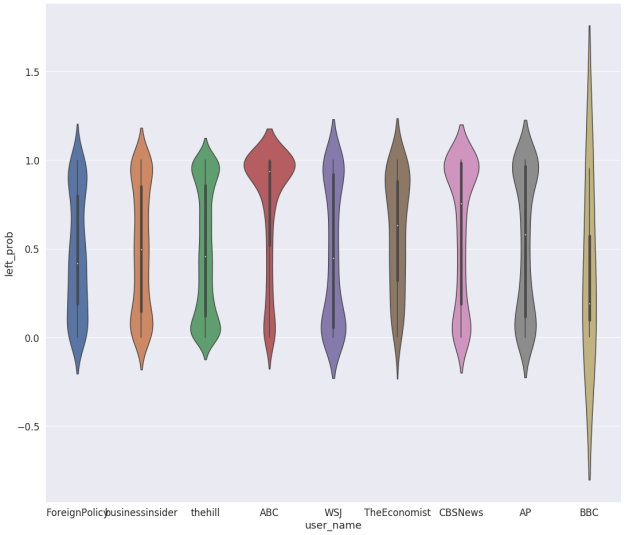
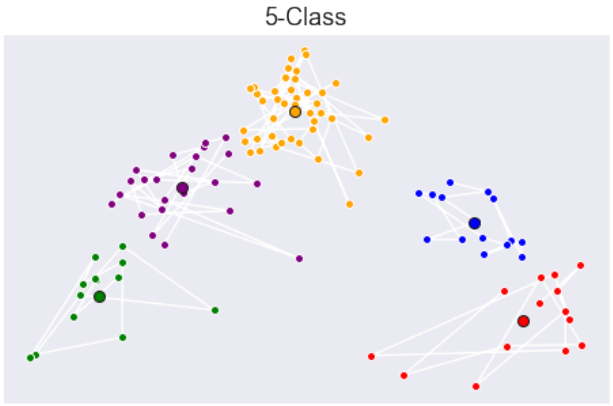


Figure 3: test_neutral

2.3 Multi-way classification

One of the major drawback of binary classification on media tweets is that there are actually more 'neutral' media compared to the left or right media. If we look at the MediaBiasChart, we can see most of the media are located on top center, which stand for neutral bias and high quality. Then the violin plot for neutral media could only be interpreted as a mixed 'left' and 'right' (See Fig 3).

On the other hand, we can split the medias into multiple groups according their bias and quality. The following clustering is computed by KMeans:



Then we study if the media in the same group are really posting similar tweets: we labeled each tweet by the group label of each source media, then split the entire dataset into 80/20 train/test and performed classification tasks.

Limited by computation power, we only tested linear SVC on the entire dataset and this model performed considerably on this 5-way classification task:

	precision	recall	f1-score	support
0	0.52	0.76	0.62	53808
1	0.71	0.52	0.60	43565
2	0.80	0.50	0.62	39715
3	0.63	0.40	0.49	45960
4	0.61	0.72	0.66	100358
avg / total	0.64	0.62	0.61	283406

3 Next Steps

We are trying to compare the multi-way classification for media and multi-way classification for tweets. For example, if we split the tweets into 5 groups according to their source media, and split the tweets into 5 groups according to their contents, would the two methods produce similar clusters? Our current result shows a large hamming distance between the result of those two methods. Part of the reason could be that the frequency of each label is severely unbalanced when we cluster the tweets according to the contents. We could try to balance the label.

Also, we are looking into the relation between NMF and KMeans clustering mentioned in [1]. So far, most of the clustering method we used are KMeans.

References

[1] Chris Ding, Xiaofeng He, Horst D Simon, and Rong Jin. On the equivalence of nonnegative matrix factorization and k-means-spectral clustering. 2008.