Datasets & Scraping

Maximilian Mozes

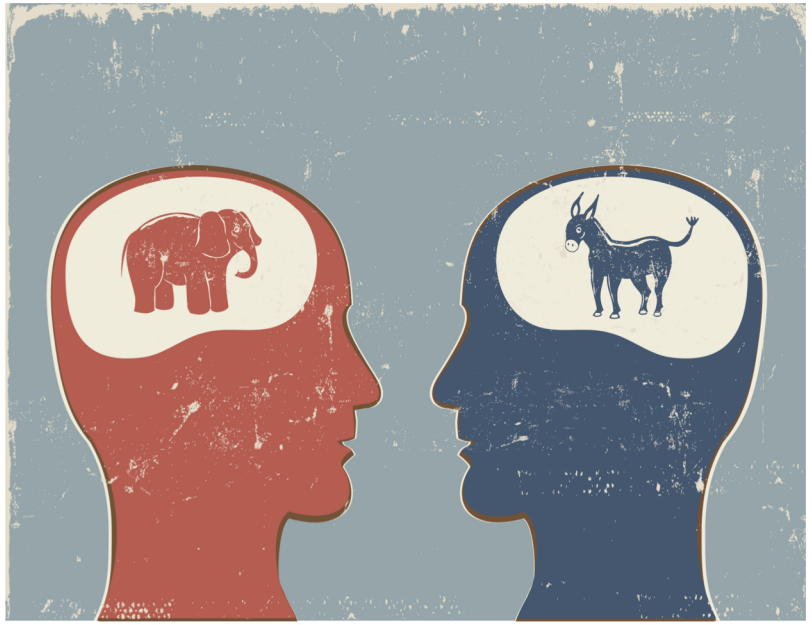
Workshop on Linguistic Temporal Trajectory Analysis

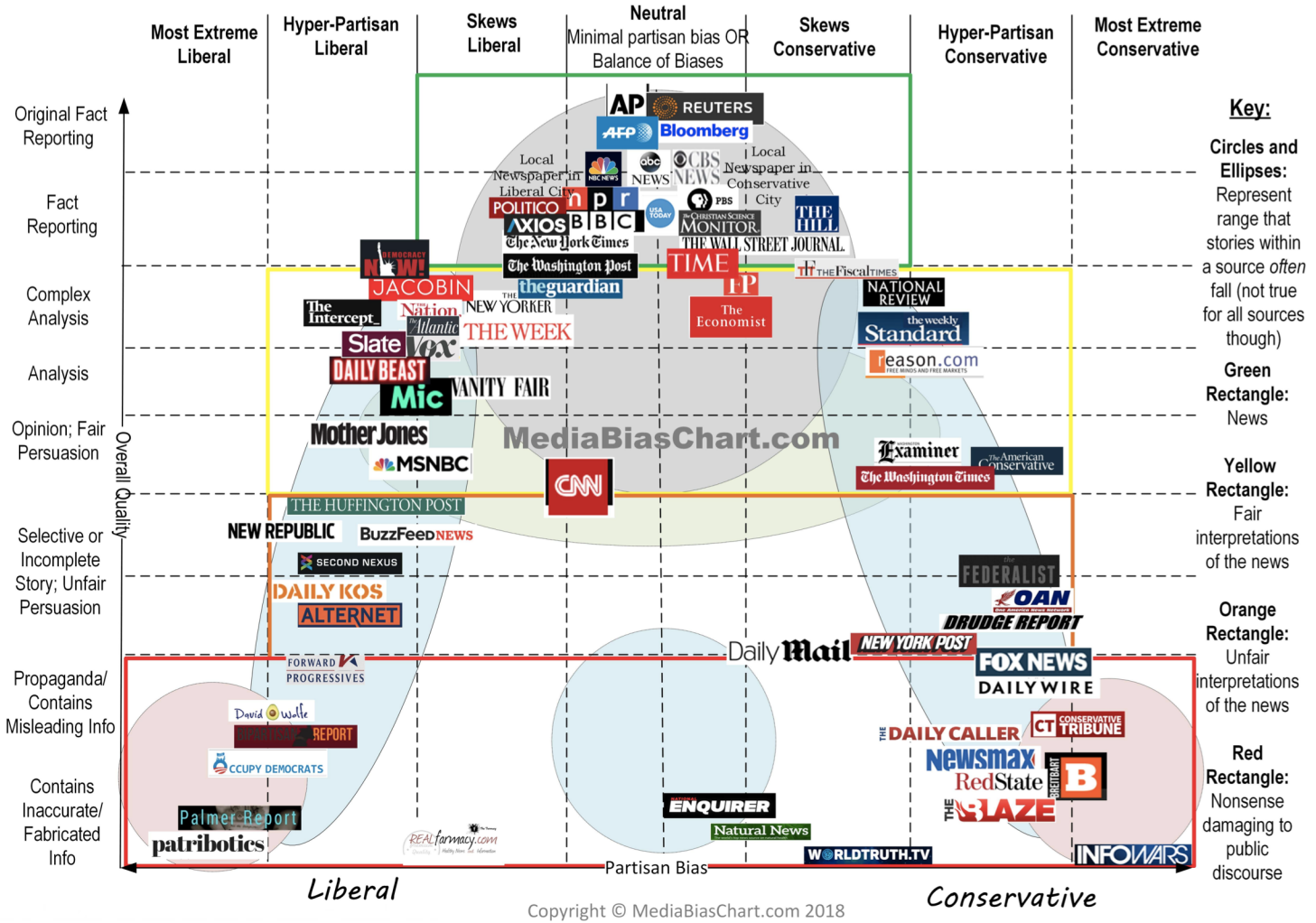
2018 European Symposium Series on Societal Challenges in Computational Social Science

December 05, 2018

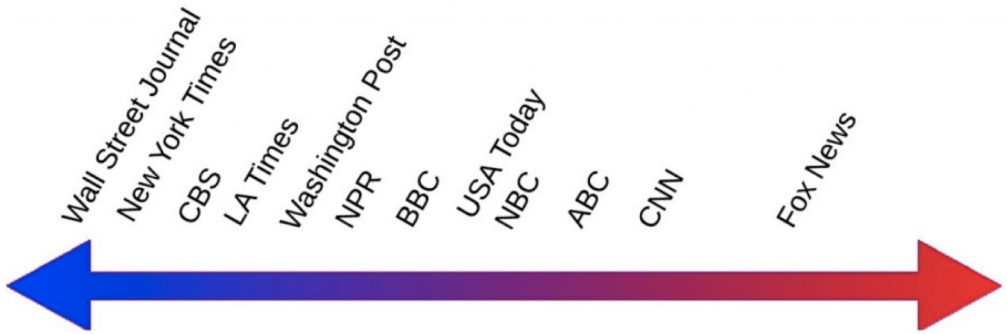
Datasets

1. Media bias



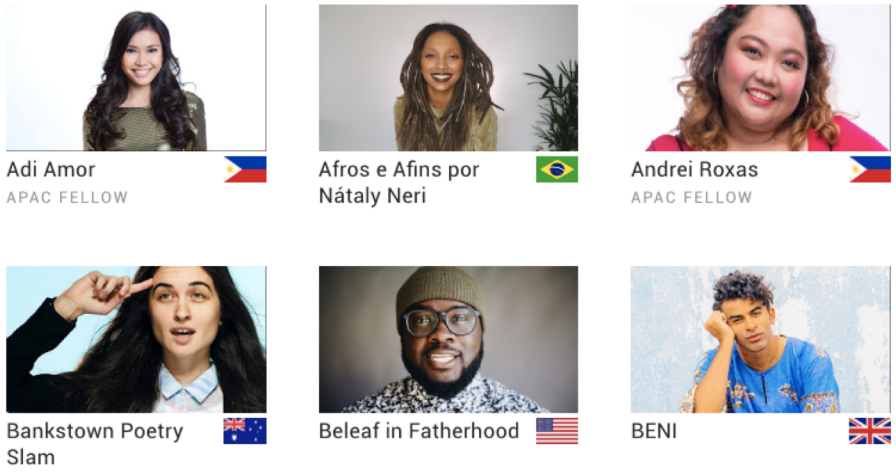


Copyright © MediaBiasChart.com 2018
or licensing and requests for royalty-free usage, please contact
mediabiaschart@gmail.com



2. YouTube Creators for Change

"Creators for Change Ambassadors and Fellows come from all over the world and use their platforms to **promote positive message** to their millions of subscribers. YouTube partnered with these inspired creators to develop video Impact Projects that aim to **bring people together** and increasing **tolerance and understanding**."



Dataset details

Political news channels

- Transcripts for videos of **8 political news channels** from YouTube
- 2000 transcripts per channel → **16000 transcripts** in total
- Attributes for each video:
  - **Transcript**
  - **Number of words**
  - **View count**
  - **Date published**
  - **Upvotes**
  - **Downvotes**

Channel selection

- We selected all English-language channels from the top 250 news channels on YouTube via **SOCIALBLADE**
- 18 news channels in total
- We matched those channels to the right/left bias ratings from **https://mediabiasfactcheck.com/right-center/**
- This resulted in 13 left and 5 right channels

Sampling

- We only consider transcripts with
  - At least 100 words
  - At least 50% of words are matched English words
  - At least 90% of words are ASCII-encoded
- We then considered only those channels with more than 2000 valid transcripts
- And randomly selected 4 left and 4 right channels
- From each channel, we randomly selected 2000 transcripts

Descriptive Statistics

Channel name	Left/right	# of videos	Mean (SD) word count	Mean (SD) view count	Total view count	Avg. # of upvotes	Avg. # of downvotes
Al Jazeera English	L	2000	1000.49 (1416.01)	20106.62 (61205.49)	40.213.246	160.64	29.12
Business Insider	L	2000	501.01 (1177.76)	120613.1 (480033.1)	241.226.161	1178.79	117.26
MSNBC	L	2000	1499.25 (906.59)	112646.1 (161002.9)	225.292.120	967.29	124.01
The Young Turks	L	2000	1504.59 (906.59)	123006.7 (96355.8)	246.013.396	2446.53	391.66
Fox News	R	2000	777.85 (506.75)	57998.42 (112583.9)	115.996.833	1094.81	108.47
Russia Today	R	2000	1420.10 (2293.65)	33572.91 (146439)	67.145.826	539.60	78.53
The Daily Wire	R	2000	4845.80 (4527.99)	79155.19 (175119.7)	158.310.378	1509.04	131.91
Rebel Media	R	2000	1250.73 (1835.69)	20356.36 (86276.11)	40.712.722	1008.68	42.07

Creators for change dataset

- Transcripts of **20838 videos** from **56 channels**
- Attributes for each video:
  - **Transcript**
  - **Number of words**
  - **View count**
  - **Date published**
  - **Upvotes**
  - **Downvotes**

Channel selection

- Selected all channels from the 2017 and 2018 **Creators for Change list from YouTube** from English speaking countries (US, UK, CA, PHI, AUS)
- Excluded music artists

Descriptive Statistics

Mean (SD) view count	**1181872 (2758495)**
Total view count	**24.627.851.276**
Mean (SD) word count	**1403.894 (1001.18)**
Mean (SD) upvotes	**34716.39 (70214.35)**
Mean (SD) downvotes	**923.60 (5055.50)**

Features (for each dataset)

- Unigrams, bigrams, trigrams
- POS proportions of all **Penn Treebank tags** (nouns, adjectives, verbs, etc.)

$$\frac{\text{\# occurrences of specific tag}}{\text{\# words}}$$

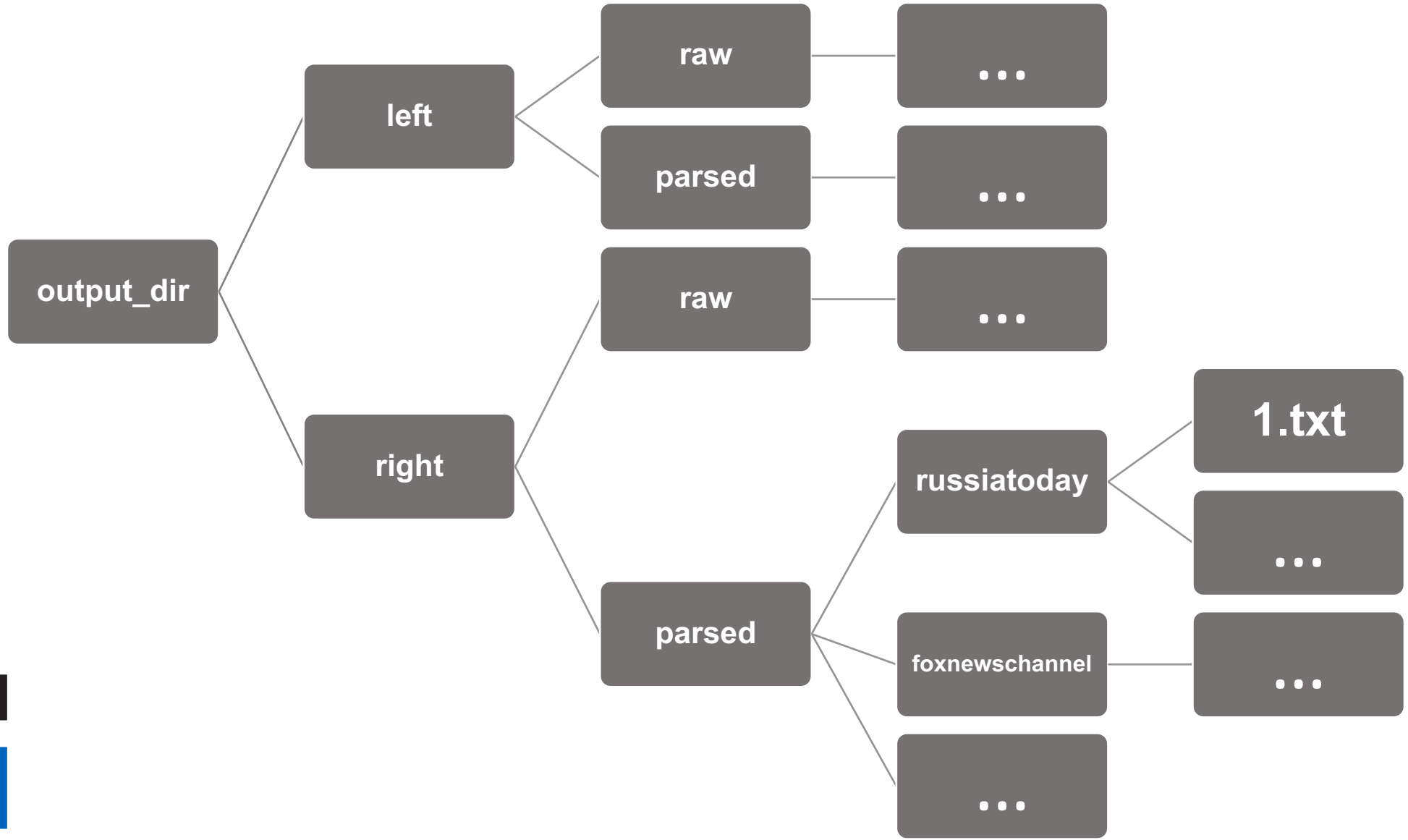
- Sentiment values with window size +/- 3, scaled to [-1,1], normalized with discrete cosine transform (low pass filter size 5)

Getting the Data

Clone **github.com/ben-aaron188/ltta_workshop**

- Political news channels in **/data_media_channels**
- Creators for change in **/ data_creators_for_change**

Hierarchy in **output_dir**



Example (output_dir/left/parsed/vicenews/1.txt)

1	I feel like I'm on a mission fighting
2	for this our entire life
3	of course I don't play identity politics
4	but it is nice to see more women in
5	office I would be the first
6	african-american elected statewide in
7	Iowa I'm a former Special Victims
8	prosecutor running for Congress I was
9	very inspired by the me to movement
10	Trump won and that changed everything
11	for me

Scraping

Retrieving the links

- We retrieved all the video URLs for a given channel using the **YouTube API v3**

Retrieving the transcripts

- We relied on **downsub.com**, a website that lets you download subtitles of YouTube videos
- Using Python's **Beautiful Soup** package, we built a web scraper that requests the transcripts for a video on downsub.com and downloads the response

Transcript decoding

1 00:00:00,000 --> 00:00:05,040 I feel like I'm on a mission fighting

2 00:00:05,040 --> 00:00:08,359 for this our entire life

3 00:00:09,710 --> 00:00:11,929 of course I don't play identity politics

4 00:00:11,929 --> 00:00:14,240 but it is nice to see more women in

5 00:00:14,240 --> 00:00:16,360 office I would be the first

6 00:00:16,360 --> 00:00:19,610 african-american elected statewide in

decoding

1	I feel like I'm on a mission fighting
2	for this our entire life
3	of course I don't play identity politics
4	but it is nice to see more women in
5	office I would be the first
6	african-american elected statewide in

Scripts available at

https://github.com/ben-aaron188/ltta_workshop/tree/master/scraping