Variation in Political News

An NLP Approach

Stephen Lee

University of Memphis

September 30, 2019

Outline

Background

"If you can't measure it, you can't improve it."

— Peter Drucker

- The 2016 United States Presidential Election raised doubts for many about the quality of their news.
- Allcott and Gentzkow [2017] estimate that the average US adult read and remembered about one, and possibly up to several, fake news articles during the election period.
- It is now not uncommon for the president of the United States to tweet that certain news sources are "Fake News".

Literature Review: Computer Science

- The LSTM architecture was introduced in Hochreiter and Schmidhuber [1997]. It builds on the Recurrant Neural Network (RNN) architecture to improve lagged information storage for the purpose of processing sequential data.
- More recently, Gers and Schmidhuber [2000], Chung et al. [2014], and Yao et al. [2015] introduce variations on the baseline Hochreiter and Schmidhuber [1997] LSTM model, although Greff et al. [2016] show them all to be roughly equivalent on a variety of prediction tasks.
- Schuster and Paliwal [1997] introduced the first bidirectional RNN reporting a significant predictive improvements over traditional RNNs. Together, bidirectional LSTM architectures prove to be among the most accurate models for language tasks, consistent with Wang et al. [2015].

Literature Review: Economics

- Gentzkow and Shapiro [2010] find that readers prefer to consume "like-minded" news and that the profit maximizing response from news companies can account for around 20% of the variation in political slant or bias.
- Gentzkow and Shapiro [2006] also find that a Bayesian consumer will reinforce their beliefs of a given news source quality when they read something that confirms their priors.
- Together, these works suggest that political news bias may tactical, and that this polarization we see may be self-reinforcing.
- Gentzkow and Shapiro [2008] go as far as to suggest that competition in information markets may actually be counterproductive in achieving balanced and unbiased news.

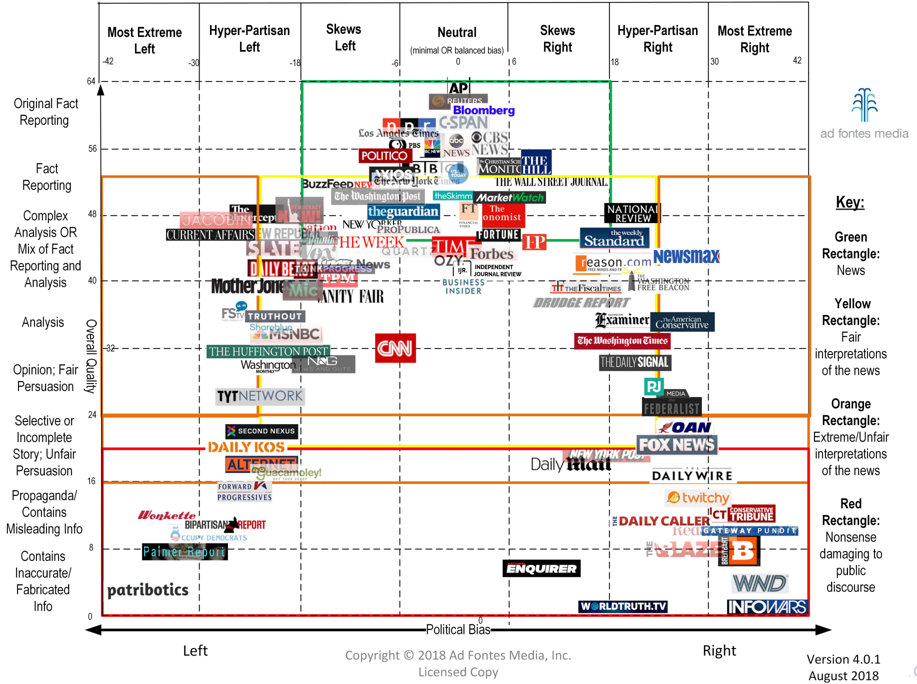
Summary

- With a goal of classifying articles by their news source, I scraped several thousand news articles from Fox, Vox, and PBS News.
- I find that a relatively simple, bidirectional, LSTM recurrent neural network can correctly predict the source of an article with very high accuracy.
- I then use the result of this trained network to build a web app that can allow for a copy-and-paste interface to interact with this classification model.
- To my knowledge, this work provides the first look at news classification using neural networks.

Data

- I mined political news articles from the websites of Fox News, Vox News, and PBS News.
- Intuitively, the motivation is to focus only on sentiment or semantics, rather than subject matter differences.
- By some estimations,[1] these three news sites represent distinct categories of news:
  - Fox as a conservative right opinion.
  - PBS as the center primary source news position.
  - Vox as a liberal left opinion.

[1] Source: https://www.adfontesmedia.com.


Most Extreme Left
Hyper-Partisan Left
Skews Left
Neutral
(minimal OR balanced bias)
Skews Right
Hyper-Partisan Right
Most Extreme Right
-42
-30
-18
-6
0
6
18
30
42
64
56
48
40
32
24
16
8
0
Original Fact Reporting
Fact Reporting
Complex Analysis OR Mix of Fact Reporting and Analysis
Analysis
Opinion; Fair Persuasion
Selective or Incomplete Story; Unfair Persuasion
Propaganda/ Contains Misleading Info
Contains Inaccurate/ Fabricated Info
Overall Quality
Political Bias
Left
Right
ad fontes media
Key:
Green Rectangle: News
Yellow Rectangle: Fair interpretations of the news
Orange Rectangle: Extreme/Unfair interpretations of the news
Red Rectangle: Nonsense damaging to public discourse
AP
REUTERS
Bloomberg
npr
C-SPAN
Los Angeles Times
PBS
abc NEWS
CBS NEWS
POLITICO
The Christian Science Monitor
THE HILL
AXIOS
BBC
USA TODAY
BuzzFeed NEWS
The New York Times
THE WALL STREET JOURNAL.
The Washington Post
theSkimm
MarketWatch
theguardian
FT
The Economist
NATIONAL REVIEW
JACOBIN
The Intercept
DEMOCRACY NOW!
CURRENT AFFAIRS
NEW REPUBLIC
The Nation
NEW YORKER
PROPUBLICA
THE WEEK
QUARTZ
TIME
FORTUNE
FP
the weekly Standard
SLATE
Vox
Forbes
OZY
reason.com
Newsmax
DAILY BEAST
THINKPROGRESS
NBC News
IJR.
INDEPENDENT JOURNAL REVIEW
THE WASHINGTON FREE BEACON
TPM
BUSINESS INSIDER
THE FISCAL TIMES
MotherJones
Mic
VANITY FAIR
DRUDGE REPORT
FSTV
TRUTHOUT
Washington Examiner
The American Conservative
Shareblue
MSNBC
The Washington Times
THE HUFFINGTON POST
CNN
Washington Monthly
N&G
NEWS AND GUTS
THE DAILY SIGNAL
PJ MEDIA
TYT NETWORK
THE FEDERALIST
SECOND NEXUS
OAN
DAILY KOS
FOX NEWS
ALTERNET
Guacamoley!
NEW YORK POST
Daily Mail
DAILY WIRE
FORWARD PROGRESSIVES
twitchy
Wonkette
CONSERVATIVE TRIBUNE
BIPARTISAN REPORT
THE DAILY CALLER
OCCUPY DEMOCRATS
RedState
GATEWAY PUNDIT
Palmer Report
THE BLAZE
BREITBART
ENQUIRER
patribotics
WND
WORLDTRUTH.TV
INFOWARS
Copyright © 2018 Ad Fontes Media, Inc.
Licensed Copy
Version 4.0.1
August 2018

Descriptive Statistics

Table: Summary statistics by news source.

Source	*Num Docs*	*Average word count*	*Pct adj*	*Pct adv*	*Average words / sentence*
Fox	661	686.2	6.6 %	3.4 %	20.1
PBS	1739	654.3	6.6 %	3.2 %	18.0
Vox	1027	1332.8	7.3 %	4.6 %	21.3

N-gram Frequencies

Table: Word Frequencies

Num	Vox	PBS	Fox
1	trump	trump	trump
2	tax	said	said
3	will	president	president
4	people	house	house
5	health	will	new
6	bill	new	will
7	republicans	white	democratic
8	one	senate	democrats
9	new	democrats	told
10	care	campaign	border

N-gram Frequencies

Table: Top frequencies of two word phrases.

Num	Vox	PBS	Fox
1	health care	white house	white house
2	white house	president donald	new york
3	trump administration	donald trump	president trump
4	donald trump	special counsel	green new
5	tax cuts	supreme court	health care
6	health insurance	attorney general	new deal
7	new york	new york	united states
8	affordable care	justice department	border security
9	tax bill	counsel robert	donald trump
10	federal government	trump said	state union

Challenges

1. Difference in corpus size from each source.
   - Bootstrap the data.
2. Variability of online formatting.
   - I removed any mention of their own organization.
   - Other common and unique affiliations.
   - Any other noticeable identifying characteristics.
3. Difference in the average article length.
   - I limited the article length to the first 500 words.

LSTM Model

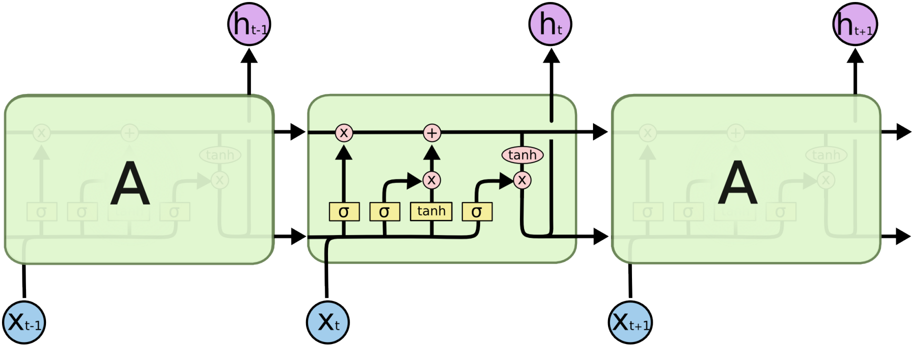


Figure: A graphical depiction of a single LSTM cell.[2]

[2]Source: https://colah.github.io/posts/2015-08-Understanding-LSTMs

Bidirectional Training

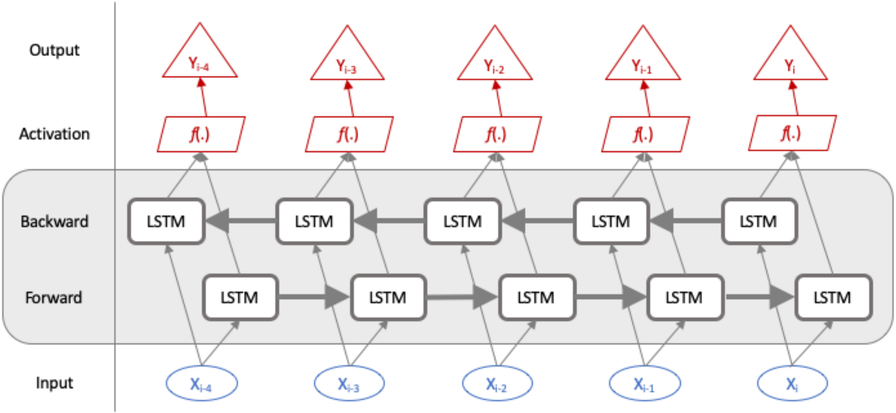


Figure: A visual representation of a bidirectional LSTM training.

Word Embedding

- I use the common crawl 840B Global Word Vector (i.e. GloVe).
- Introduced in Pennington et al. [2014] and uses 840 billion tokens and a case-sensitive vocabulary of 2.2 million words to map words into a corresponding 300×1 dimensional vector.
- Accordingly, I do minimal preprocessing to the text besides the basic cleaning mentioned previously.

Methodology

- I split the data into training (90%) and testing (10%).
- Using the training data, I fit a birdirectional LSTM using a range of parameterizations, and then calculate F1 scores using the unseen testing dataset.
- Similarly, I compare the bidirectional model to a forward only model using the same approach.

Note,

$$F_1 = 2\ \frac{Precision \cdot Recall}{Precision + Recall}$$

Where,

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

Results: Bidirectional LSTM

Table: Training results from the bidirectional LSTM, sorted by F1 score.

Article Length	Batch Size	Dropout	Recurrent Dropout	Steps Per Epoch	F1
250	64	0.1	0.2	1000	0.946
500	64	0.2	0.2	1000	0.944
500	64	0.2	0.1	1000	0.939
250	64	0.1	0.1	1000	0.937
500	64	0.1	0.1	1000	0.937
500	64	0.1	0.2	1000	0.933
250	64	0.2	0.1	1000	0.921
250	32	0.2	0.1	1000	0.910
250	32	0.1	0.1	1000	0.906
...	...	...	...	...	...
500	32	0.1	0.1	500	0.828

Results: Forward LSTM

Table: Training results from the unidirectional LSTM, sorted by F1 score.

Article Length	Batch Size	Dropout	Recurrent Dropout	Steps Per Epoch	F1
250	64	0.2	0.1	1000	0.824
250	64	0.1	0.2	1000	0.797
250	32	0.1	0.2	1000	0.766
500	64	0.1	0.1	1000	0.724
250	64	0.2	0.2	1000	0.716
250	32	0.2	0.1	1000	0.703
500	64	0.2	0.2	1000	0.703
250	32	0.2	0.2	1000	0.695
250	64	0.1	0.2	500	0.686
...	...	...	...	...	...
500	64	0.2	0.1	500	0.556

Demo

Figure: The web app;s home page.

Conclusion and Discussion

- I've shown that NLP techniques can accurately classify news articles based on language differences in the underlying news sources.
- Given Gentzkow and Shapiro [2008] and Gentzkow and Shapiro [2006], it seems unlikely that biased news (or even fake news) will disappear anytime soon.
- A web application based on these ideas could serve as a starting point to measure news consumption. Analogus to:
  - Calendars can help to measure our time use.
  - Nutrition apps measure our macronutrients.
  - GPS measures our geographical position.
  - Can we have an app to measure our news consumption?

End

Thank you for your time.
Questions?

References

Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36, 2017.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Matthew Gentzkow and Jesse M Shapiro. Media bias and reputation. *Journal of political Economy*, 114(2):280–316, 2006.

Matthew Gentzkow and Jesse M Shapiro. Competition and truth in the market for news. *Journal of Economic perspectives*, 22(2):133–154, 2008.

Matthew Gentzkow and Jesse M Shapiro. What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71, 2010.

Felix A Gers and Jürgen Schmidhuber. Recurrent nets that time and count. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, volume 3, pages 189–194. IEEE, 2000.