Predicting Media Bias in Online News
CS 229: Machine Learning - Final Project

John Merriman Sholar (jmsholar@stanford.edu) & Noa Glaser (SuNet ID: noaglasr@stanford.edu)

June 6th, 2016

Abstract

This paper explores applications of machine learning to analyzing media bias. We seek patterns in event coverage and headlines across different news sources. For headline wording, we first validate the existence of informative trends by examining the performance of multinomial Naive Bayes and SVM classification in mapping titles to news sources. We then perform keyword analysis to determine which words are most indicative of certain news sources. In event coverage, we use unsupervised clustering techniques to profiles news sources by the events covered. We vary the scope of our analysis from global news to Israel and Palestine from 2014 to 2016 and Israel during the summer of 2014. We were able to observe meaningful trends in both headline key words and event coverage and are excited about this methodology as an objective lens to the analysis of media bias.

1 Problem and Background

In cognitive science, bias is defined as deviation from the norm, or true value [1]. Media bias can refer to deviating coverage amounts across event types or skewed representation of the events. Because news sources have authority and influence over popular opinion, this bias incredibly important to monitor.

Previous work has examined geographical overreporting, variation in event coverage promptness, differences in writing-style readability, and variation in intensity of coverage. Other work examines biased adjectives and utilizes natural language processing to understand bias in writing style. Mladenic examines networks of cross referencing between news sources and news providers to understand which voices news sources are choosing to represent. We adopt a Naive Bayes model for keyword analysis - discerning the words most indicative of which source is reporting about a certain topic. For example, Leban used keyword analysis to study bias across a variety of subjects, including the conflict in Crimea.

Media bias affects all stages of news publishing. Because headlines most affect the general consumer, we focus on wording and event selection (cherry picking or selection bias).

2 Data

For the data for this project we use the eventregistry.org API [5].Event Registry [ER] collects news articles from RSS feeds of over 100,000 news sources around the world. ER also clusters groups of 'articles' into 'events' based on location and article content. These ER clusters will be referred to as 'event' in the rest of this paper.

Most data analysis was conducted with the SciKit-Learn Python machine learning framework. [6]

2.1 Headlines

For an initial phase of the project, we used the Event Registry API to curate over 160,000 article headlines for articles published by the top twenty news websites (as ranked by Alexa web traffic metrics) between 2014 and 2016. The results of applying keyword analysis to this dataset were used as a baseline for our main goal of applying a similar analysis to media surrounding the Israel-Palestine conflict.

For the second phase of the project, we used the Event Registry API to curate over 1,500 articles, focused specifically on the Israel-Palestine conflict. For this dataset, 8 news sources were selected specifically for their collective propensity to provide a wide range of opinions on the conflict.

For both the baseline and primary datasets, Naive Bayes and SVM models were trained to predict the news organization that published an article, given the headline of the article. Article headlines were preprocessed using a combination of SciKit-Learn's Count Vectorizer and TF-IDF Transfomer tools.

2.2 Events

To study event selection bias, we gathered data for 600 events related to Israel between June 1st and September 30th, 2014. The data included 1,143 news sources; all news sources were used to normalize vector norms but only the top 100 (by total number of articles) were clustered.

3 Methodology

3.1 Headlines

Multiclass Naive Bayes and SVM models were trained on the dataset described in section 2, attempting to predict the news organization that published a given article based on the headline of that article. Accuracy of these classifiers was used as an indicator of the feasibility of pursuing keyword analysis (under the hypothesis that the existence of observable trends in data would lend itself to worthwhile results under keyword analysis). Accuracy statistics can be found in section 4.1.

Having verified the existence of observable trends in data, we generate for each unique pairing of token and news organization a measure of "indicativeness", or how representative the given token is of article headlines produced by a given news organization. We note that the Naive Bayesian Model generates probabilities of the form $P(\text{token} \mid \text{news outlet})$. Using these, we can calculate indicativeness for each pairing of token and news outlet:

$$\text{Indicativeness} = \log\left(\frac{P(\text{token} \mid \text{news outlet})}{P(\text{token} \mid \text{NOT news outlet})}\right)$$

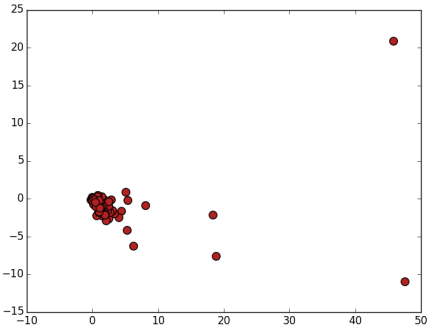
A summary of the most indicative keywords for each news organization can also be found in section 4.1.

3.2 Events

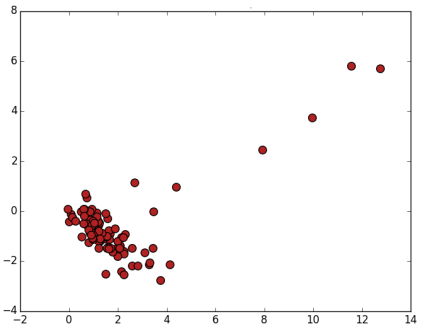
We tested the hypothesis that there exists systematic event selection bias which would allow us to create meaningful profiles of news sources.

We examined three models for news sources: coverage propensity - vector of number of articles covering each event, event cherry picking - vector of binaries indicating whether each event was covered, and normalized propensity vectors. A PCA plotting the 100 most common news sources under the three models is presented in Figure 1.

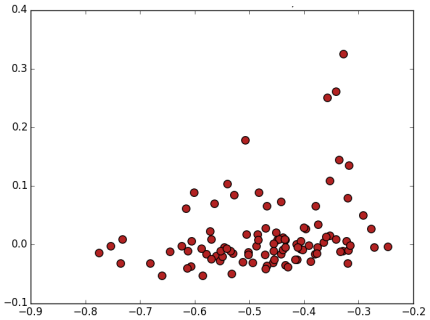
Figure 1: Visualization of the three news source models [Left: number articles/event, Center: Whether reported on event, Right: normalized number of articles/event]. PCA of 600 event dimensions to 2D. Plotted are 100 news sources with most articles about the Israel in the summer of 2014.



(a) Propensity model represents local news as strong outliers.



(b) Binary model has more spread, local news still separate



(c) Normalized propensity

Gaussian Mixture Models and Hierarchical Clustering Models resulted in similar outlet profiles and so we proceeded with KNN.

4 Results

4.1 Headlines Keyword Analysis

Accuracy statistics achieved by the Multinomial Naive Bayes and SVM classifiers on the larger international news dataset (160,000 articles) are reported below.

Model	Precision	Recall	F_1 Score
Naive Bayes	.50	.40	.39
SVM	.47	.48	.48

As was noted in section 2, results on the larger dataset were intended to act as a baseline for the accuracy of these same classifiers when applied to the smaller, more focused dataset of articles covering the Israel-Palestine conflict. The results of classification on this dataset are presented below.

Model	Precision	Recall	F_1 Score
Naive Bayes	.57	.53	.52
SVM	.47	.48	.48

For the baseline dataset, we present the most indicative tokens for each news organization. The existence of observable, sensible trends lends confidence to the corresponding predicative keywords for the Israel-Palestine dataset.

News Organization	Most Indicative Tokens
CNN	cnn, com, cnnpolitics, isis, facts, 370, opinion, mh370, plague, cruz
Bloomberg	bloomberg, said, draghi, yuan, treasuries, estimates, pboc, bonds, ruble, traders
Huffington Post	huffington, jenner, here, post, yoga, kardashian, these, this, thing, adorable
BBC	bbc, news, edinburgh, glasgow, ni, utd, lorry, labour, wales, belfast

For the primary dataset we found the following most indicative keywords for each news outlet:

News Organization	Traditional Reputation	Most Indicative Tokens
Fox News	Conservative American	fox, claim, site, holy, nations, western, muslim, prepares, down, holocaust
Reuters	Moderate International	treaty, solidifying, reuters, update, kill, vatican, agrees, relationship, troops, first
Haaretz	Liberal Israeli	jewish, haaretz, bid, live, watch, world, lawmakers, probe, 2016, vote
Jerusalem Post	Moderate Israeli	zionism, encountering, german, one, india, process, fate, candidly, working, that
Israel Hayom	Conservative Israeli	hayom, israel, turkey, jews, mind, caught, european, hamas, blackmail, pm
Arutz Sheva	Conservative Israeli	global, agenda, news, part, inside, middle, swedish, time, east, internet
Palestine News Agency (WAFA)	Liberal Palestinian	newspaper, review, dailies, focus, newspapers, highlight, killing, premier, international, rome
Palestine Chronicle	Conservative Palestinian	chronicle, palestine, book, apartheid, nakba, the, zionist, media, bds, struggle

While we included "traditional reputation" for each news source for context, these are understandably subjective qualities and do not reflect the opinions of the writers. These labels reflect what we believe to be general public opinion.

4.2 Events

We polled 22 Stanford students with varying degrees of familiarity with the news sources and events to rank 9 KNN clusters of news sources based on the intergroup coherency and insight. Each proposed clustering was given a score from 1 to 10. The results, normalized per respondent mean and variance, are shown in the table below.

	Fewer clusters	More Clusters
Event Frequency	KNN4: -0.1493	KNN8: 0.0163
Event Binomials	KNN3: 0.5005 KNN4: 0.4173	KNN8: -0.1171
Normalized event frequency	KNN3: 0.2558 KNN4: -0.0708	KNN6: 0.7173 KNN8: 0.3866

Article frequency based clustering was quite unpopular as it clustered local news outliers into very small clusters and lumped together the remaining sources. This behavior also emerged, to a lesser extent, with event binomials. Event binomials with fewer clusters and normalized vectors with more clusters were the most popular.

Respondents most preferred clustering normalized event vectors into 6 groups, which produces the following:

1. The Jerusalem Post, Arutz Sheva, Haaretz.com, www.israelhayom.com, ynet, The New York Times, Jewish Journal, The Washington Post, The Guardian, The Independent, Algemeiner.com
2. TIME, Los Angeles Times, Truthdig, DIE WELT, The Japan Times, San Francisco Gate, Ad Hoc News, The National, N24.de, Thomson Reuters Foundation, The Irish Times, www.greenpeace-magazin.de, US News & World Report, www.americanthinker.com, The Sydney Morning Herald, USA Today, www.france24.com, POLITICO, The Huffington Post UK, europenews.dk, The Inquisitr News, ABC News, Business Insider, The PJ Tatler, www.montrealgazette.com, LaVanguardia, DailyTimes, www.middle-east-online.com
3. Yahoo News, uDW.com, Independent.ie, Economic Times, The Wall Street Journal, Sify, El Economista, Daily News and Analysis (DNA) India, Arab News, The Indian Express, GlobalSecurity.org, El Economista

(EcoDiario), The Sacramento Bee

4. www.presstv.ir, Irish Sun, International Business Times UK, www.aa.com.tr, www.albawaba.com, www.palestine-info.co.uk, Naharnet, News From Antiwar.com, english.wafa.ps
5. BBC News, NDTV, CBC News, The Globe and Mail, Telegraph.co.uk, Xinhuanet.com, Reuters, Mail Online, Miami Herald, Fox News, news24, The Charlotte Observer, The Christian Science Monitor, VOA Voice of America, NPR.org, euronews, Philly.com, Bloomberg Business, Boston Herald, Channel NewsAsia, The Hindu, theStar.com, Daily News, Zee News, Manila Bulletin, National Post, timesofmalta.com
6. GULF NEWS, The Huffington Post, The Daily Star Lebanon, ABC News, CNN International, english.farsnews.com, Republika Online, Star Tribune, ReliefWeb, Sky News, The Star Online, RT

5 Conclusions

We believe that word counts and event coverage profiling can serve as a highly objective lens for the study of media bias. Common approaches in NLP, such as quantifying inflammatory adjectives or the readability of text undoubtedly introduce bias and are hard to generalize across new languages. This type of analysis could hold news sources more accountable than one fraught with subjective metrics.

The results of headline keyword analysis proved particularly interesting, and we observe that the keywords judged to be most indicative of various news outlets display significant correlation with the established political leanings of each outlet. Additionally, the results seen here prompt new and more exciting questions and applications surrounding this research. Several immediately apparent next steps include a rigorous evaluation of the idea of "indicativeness" (and an analysis how best to compute this metric), an expansion and cleansing of the dataset (which was subject to the limitations of the EventRegistry API), and an exploration of practical applications of these trends.

Interesting trends emerged in events clustering, although the model is quite naive. For example, the clustering on page 3 groups newspapers stereotypically geared towards Israelies and Jewish Americans (cluster 1), Palestinian and Irish/British sources (cluster 4) and liberal German and American news sources (cluster 2). We believe that more meaningful clusters can be generated by adding features such as event categories/keywords.

Much of the inspiration for this research came as a result of the authors' own experiences with the so-called "echo-chamber effect", in which an individual consumes only media that validate his or her views. In attempting to classify news outlets based on their political leanings and biases, one potential application of this research would be to generate a set of news sources representing a comprehensive span of opinions on a given issue. Such an application would hopefully promote a greater awareness of the intricacies of important issues, and facililate a more objective, productive discussion surrounding them.

References

[1] Mladenic, Dunja. "Learning How to Detect News Bias." (n.d.): n. pag. 2015. Web. 21 May 2016.

[2] Flaounas, Ilias. "Pattern Analysis of News Media Content" Diss. U of Bristol, 2011. Print.

[3] Flaounas, Ilias and Omar, Ali and LAnsdall-welfare, Thomas, etall, "Research Methods In the Age of Digital Journalism." Digital Journalism, 102-116. 2013.

[4] Leban, Gregor. "News reporting bias detection prototype".www.xlike.org

[5] Leban, Gregor, Bla Fortuna, Marko Grobelnik, Bla Novak, and Alja Komerlj. "Event Registry." Event Registry. N.p., n.d. Web. 23 May 2016.

[6] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[7] "Alexa Top News Sites." Alexa.com. Amazon, n.d. Web. 22 May 2016.