Reading Between (the Party) Lines

by

Sophie Beiying Chou

Submitted to the MIT Media Lab,
School of Architecture and Planning
in partial fulfillment of the requirements for the degree of

MS in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2016

© Massachusetts Institute of Technology 2016. All rights reserved.

Author ....
MIT Media Lab
May 5, 2016

Certified by ....
Deb Roy
Associate Professor
Thesis Supervisor

Accepted by ....
WHO IS THE CHAIR(WO)MAN?
Chairman, Department Committee on Graduate Theses

Reading Between (the Party) Lines

by

Sophie Beiying Chou

Submitted to the MIT Media Lab
on May 5, 2016, in partial fulfillment of the
requirements for the degree of
MS in Media Arts and Sciences

Abstract

TO-DO

Thesis Supervisor: Deb Roy
Title: Associate Professor

The following people served as readers for this thesis:

Sepandar Kamvar. . . . .

Associate Professor of Media Arts and Sciences

MIT Media Lab

Iyad Rahwan . . . .

Associate Professor of Media Arts and Sciences

MIT Media Lab

Acknowledgments

Thank you !!

Contents

Chapter 1

Introduction

Most Americans say that they want to read news that's unbiased. A survey from Pew Research in 2012 showed that more than two-thirds (68%) of readers want to read political articles with a neutral stance, compared to just a little less than a quarter (23%) of those who want to read those stories that share their point of view.1 But what exactly does that mean?

To begin with, whether or not we perceive news as biased is biased in itself. Conservative readers tend to view media as more biased than both Democrats and Independents (49% to 32% and 35%, respectively)[**?**].

The Hostile Media Effect, first studied by Vallone, Ross, and Lepper in 1985, gives one possible explanation for discrepancies: it describes a phenomenon where people with strong stances on an issue tend to perceive media covered as biased against their opinions, even on the same article.2

Clearly, finding bias in news depends on who the reader is as much as what they are reading.

In my thesis, I seek to examine the effects of context versus content in perceptions of media bias. In particular, when the context of a story is removed, how do linguistic features, in particular reading level and vocabulary, in the content affect the reader? Although studies have been conducted to both examine the psychological effect of wording on believability (see “Seductive Allure”) and the impact of media brands and bias (see Baum, 2008), I seek to combine and contrast the two.

To do so, I will perform an A/B study for a broad range of readers to read and annotate political news stories (collected daily and sorted using a machine learning classifier). Each story is determined to be primarily about one political candidate and one topic computationally. In the control group, readers are given the full text of the article with no additional content. In the experimental group, readers are given a link to the original article complete with the byline, publication, and images. Stories are classified as either “high reading level,” “average reading level,” or “low reading level” by the Flesch-Kincaid test.

For each reader, I will collect their demographic information, and self-reported political stances. I will then analyze the effects of reading level versus media brand in the reader’s perception of the article.

I want to measure just how strong the effect of the media brand and the reader’s beliefs are.

Chapter 2

The Power of (Percieved) Media Bias

2.1 The Effects of Media Bias

Why is media bias IMPORTANT? Why is the problem IMPORTANT?

- Fox News Effect - Does the media matter

2.2 The Role of the Reader in Perceptions of Bias

It comes as no surprise that our own political stances have a significant effect in our perceptions of bias in the media.

In even seemingly neutral stories, partisans tend to view reporting as biased against their own views. This phenomenon– deemed the "hostile media effect"– was first studied at Stanford University by Robert P. Vallone, Lee Ross, and Mark R. Lepper in 1985 [4]. Although "true" neutrality of a story is nearly impossible to quantify due to the subjective nature of the concept, Vallone et. al were able to successfully demonstrate that partisans of *both* sides (pro-Israeli and pro-Arab) viewed the same news segments as hostile towards their beliefs and favorable to the other side.

? Perceptions of media bias, then, have as much to do as self-serving motivations to secure preferential treatment as they do with the media itself.

? The political leanings of the reader are essential considerations when attempting to measure other factors that contribute to bias. In

2.3 The Role of Media Brands in Perceptions of Bias

The media, of course, is not just one unified mass, and in an increasingly fragmented ecosystem, the role of media brands is a crucial factor in the perception of bias. Although most research [1]

For instance, most research on the hostile media phenomenon conceptualizes the news media as an undifferentiated mass of information sources that individuals can (and do) reasonably characterize as having a uniform political orientation (Giner-Sorolla and Chaiken 1994, Peffley et al. 2001, Eveland and Shah 2003). Yet, the past two decades have seen a dramatic increase in the number and variety of news sources. One consequence is that Democrats and Republicans are increasingly likely to differ systematically in their assessments of specific media outlets.

With the decline of print newspapers, a diverse number of new platforms and web-centric publications have risen.

How do you control for the above things in your study?

2.4 The Role of Language [Policial Persuasion]

2.4.1 Language and Politics

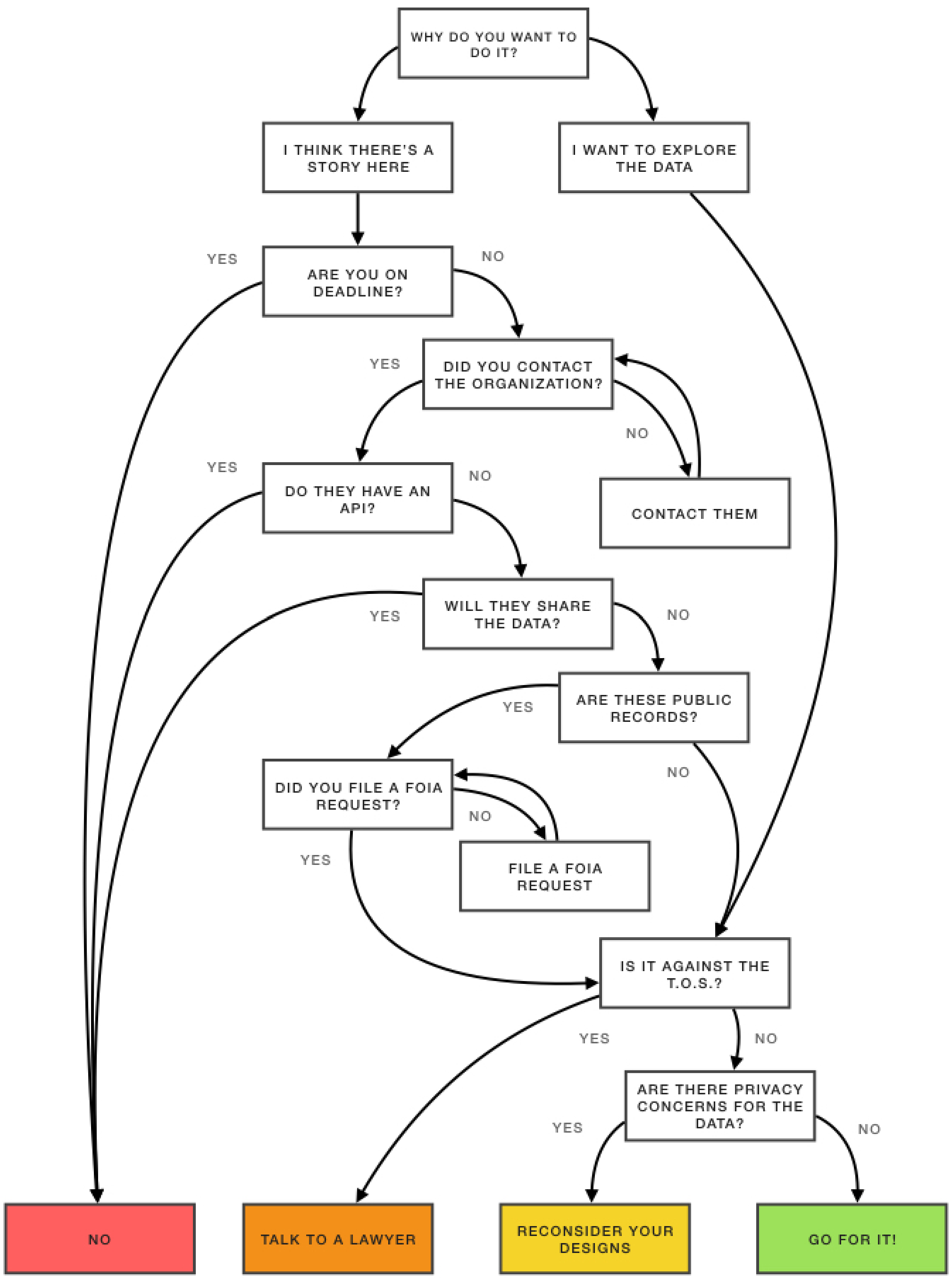
Presidential speeches degrading over time– ie simple language appeals to the masses in politics

2.4.2 The Seductive Allure [... of Simple] Language

But we trust complex language for explaining technical facts

Test image

Should You Build a Scraper?



2.5 Importance in Political Outcomes

Fox news effect

2.6 The 2016 Elections

2.6.1 Criticism of Media Bias

(Obama Speech)

So.... are you what you cover?

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 3

Data Collection

3.1 The Electome

The Electome is a large, collaborative, and ongoing effort in the Laboratory for Social Machines that seeks to analyze the "competition of ideas" in the upcoming 2016 elections. It does so by using techniques in natural language processing, machine learning, and network analysis to make sense of "big data" collected from two main sources: traditional media (online versions of news publications) and social media (Twitter) [5].

The foundations of this thesis, which emerged from the Electome, are grounded in the former dataset, although only a portion of the data collected is analyzed in this study.

3.2 Story Collection

News articles from 14 different news publications were systematically collected every hour from RSS feeds beginning from January 2015. The outlets tracked are:

1. CNN
2. Fox News
3. The Wall Street Journal
4. ProPublica
5. Politico
6. McClatchy
7. The Washington Post
8. Buzzfeed (News only)
9. National Public Radio (NPR)
10. The Huffington Post
11. The Associated Press
12. Reuters
13. The New York Times
14. The Los Angeles Times

The above outlets were chosen to form a diverse subset of the current U.S. news ecosytem, including a combination of private and public, liberal and conservative, legacy and new media publications. Also included are wire services and a mix of media delivery formats for which the outlet is known (radio, television, print, or web).

Steps to collect the news stories were as follows:

1. For each news publication:
    (a) Use regular expressions to extract all RSS feed urls for a news site.
    (b) For each RSS feed:
        i. Parse feed using open source xml reader library, Feedparser.
        ii. For each link to a story in the feed:
            A. Parse html using Beautiful Soup 3 (an open source python library)
            B. Insert headline, authors, story text, publication date and retrieval date into an SQL database.

Data depulication (by story url and headline) is then performed to ensure only one copy of each article is in the database. This step is necessary as articles from wire services often appear across many outlets and effect aggregate text analsyis.

On average, 2,000 stories are collected per day across all outlets. However, volume follows a consistent pattern of fluctuation depending on weekday, ranging from approximately 1,000 to 3,000 stories.

[INSERT HERE GRAPH OF NEWS STORIES VOLUME BY WEEKDAY]

As of March 1st, 2016, there were 855,000 stories collected in the database and 43,000 journalists.

For the purposes of this study, stories were examined from five outlets:

1. CNN

2. Fox News

3. The New York Times

4. The Wall Street Journal

5. The Associated Press

because: and cite Pew

3.3 Election Classification

[How do I cite Prashanth's unpublished work?]

3.4 Article Topic Classification

[How do I cite Prashanth's unpublished work?]

Election Classifier The election classifier is a binary classifier which takes a news article as input and determines whether it is about the 2016 US election or not. Since news articles usually con- tain clean and structured language, they can easily be classi- fied as election-related using Bag-of-Word (BoWs) features. We used the chi-square test for feature selection. Chi-square measures the lack of independence between a term in an ar- ticle and a class (in this case the election). High scores on chi-square indicate that the null hypothesis of independence should be rejected and thus that the occurrence of the term and class are dependent. The features are ranked based on their scores and the top 20,000 features form the vocabulary for the binary classifier. Next, using scikit-learn (Pedregosa et al. 2011)âĂŞa Python machine learning libraryâĂŞ a binary Maximum Entropy (MaxEnt) text classifier (Nigam, Laf- ferty, and McCallum 1999) is trained on a balanced dataset of 1,000 manually labelled news articles. The classifier was evaluated on a separate balanced test set of 300 articles, with

- Income Inequality
- Environment/Energy
- Jobs/Employment
- Guns
- Racial Issues
- Foreign Policy/National Security
- LGBT Issues
- Ethics
- Education
- Financial Regulation
- Budget/Taxation
- Veterans
- Campaign Finance
- Surveillance/Privacy
- Drugs
- Justice
- Abortion
- Immigration
- Trade
- Health Care
- Economy
- Other

3.5 Flesch-Kincaid Readability Tests

In this study, we focus primarily on the Flesch-Kincaid (F-K) tests for estimating text readability. Originally developed for the U.S. Navy in 1975 for assessing the difficulty of technical manuals, the F-K reading level corresponds roughly to U.S. grade level and the reading ease score is inversely proportional to the grade level on a scale from 0 to approximately 120 [2].

We chose the F-K tests over other comparable ones due to its popularity in educational assessment and other applications, including in legislation. For example, it is required by law in Florida that life insurance policies have a Flesch reading ease of 45 or greater (less than 12th grade in reading level) [3]. The F-K tests are also bundled in many common word processing services, including Microsoft Office Word. As a comparison, basic article analysis is also computed using the Gunning fog index (see Section 5.2.1).

The formula for Flesch reading ease is as follows:

$$206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

And for reading grade level:

$$0.39\left(\frac{\text{total words}}{\text{total sentences}}\right) + 11.8\left(\frac{\text{total syllables}}{\text{total words}}\right) - 15.59$$

The two formulas are not directly comparable due to the difference in weighting factors. For ease of metaphor, we use the grade level tests in our analysis. Syllable length is highly weighted in this formula, so it is possible to generate a story of very high reading level that consists of a single word in a single sentence (the longest English word, *pneumonoultramicroscopicsilicovolcanoconiosi*, a type of lung disease, has a reading grade level of 197.2), which is a limitation of the method, since texts with polysyllabic words are not always necessarily more difficult to read.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 4

Experimental Design

4.1 Data Selection

4.2 CrowdFlower

4.3 Demographic Survey

4.4 Political Affiliation Survey

4.5 Quality Assurance

-Filter by nationality - highest setting on crowdflower - Gold questions - time limits - price

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 5

Pre-Survey Analysis

5.1 Topic Analysis

5.2 Flesch-Kincaid Analysis

5.2.1 Comparisons to other Reading Level Tests

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 6

Study

We ran this over n days blah blah

6.1 Demographics of Readers

6.2 Overall Bias Reportings

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 7

Analysis

7.1 Media Brand Effect

7.2 Reading Level Effect

7.3 Other Linguistic Cues

THIS PAGE INTENTIONALLY LEFT BLANK

Bibliography

[1] Matthew A Baum, Phil Gussin, et al. In the eye of the beholder: How information shortcuts shape individual perceptions of bias in the media. *Quarterly Journal of political science*, 3(1):1–31, 2008.

[2] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, DTIC Document, 1975.

[3] The Florida Legislature. The 2015 florida statutes title xxxvii. `http://www.leg.state.fl.us`. (Accessed on 04/05/2016).

[4] Robert P Vallone, Lee Ross, and Mark R Lepper. The hostile media phenomenon: biased perception and perceptions of media bias in coverage of the beirut massacre. *Journal of personality and social psychology*, 49(3):577, 1985.

[5] Prashanth Vijayaraghavan, Soroush Vosoughi, and Deb Roy. Automatic detection and categorization of election-related tweets. In *10th International AAAI Conference on Web and Social Media*, 2016.