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

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

Lorem ipsum dolor sit amet, no nibh deleniti pri, docendi omnesque no cum, sed rationibus consetetur ne. Nam mentitum maluisset te, est eleifend intellegebat ex. Stet volutpat deseruisse pro an, at causae alienum assueverit vel. Vis timeam atomorum cu, solet epicurei temporibus ut ius. Pertinax consetetur sea te. Ne quas harum denique ius. Et sit vocibus sententiae definiebas, ei usu minim abhorreant. Nam cu errem equidem, omnesque offendit ea duo. Duo an dicant definitiones. Tation graece melius cum ut, ea dicta vulputate reprehendunt vix, eu quis fuisset expetendis mea. Has blandit praesent reprehendunt ei. Animal iuvaret has ea, vis quodsi sanctus an. Duo albucius hendrerit definitionem at, vide malorum vel an. No sit debet blandit, mentitum temporibus cu sea. Id vitae aperiam vis, virtute copiosae accusata no ius. Invenire dignissim at cum, an adhuc vivendo principes has. Ut mei mutat voluptua suavitate, aliquid equidem has et. Cum eu erant putant, ne facete timeam euismod sed, usu ei erroribus hendrerit. Est id vero dictas legendos. Et ullum iriure mel, ei eum graeci interpretaris, pro atqui oblique id. Enim mundi liberavisse mel ei, pri et quodsi eleifend. Habeo molestie quo et, mundi primis accumsan eu vim, pro ei impetus prodesset efficiantur.

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

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

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

Content vs. Context in Percieved Media Bias

2.1 The Role of the Reader in Percieved Bias

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

2.1.1 The Hostile Media Effect

Test test [1]

2.1.2 Perceptions of Media Brands

2.2 The Role of Language in Percieved Bias

2.2.1 Language and Politics

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

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

But we trust complex language for explaining technical facts

2.3 The 2016 Elections

2.3.1 Criticism of Media Bias

(Obama Speech)

So.... are you what you cover?

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) [2].

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:

- CNN
- Fox News
- The Wall Street Journal

- ProPublica
- Politico
- McClatchy
- The Washington Post
- Buzzfeed (News only)
- NPR
- The Huffington Post
- The Associated Press
- Reuters
- The New York Times
- 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:

- CNN
- Fox News
- The New York Times
- The Wall Street Journal
- The Associated Press

because: and cite Pew

3.3 Election Classification

Acknowledge Prashanth

3.4 Article Topic Classification

(This part is prashanths– so cite)

- 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

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

Appendix A

Tables

Table A.1: Armadillos

Armadillos	are
our	friends

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix B

Figures

Figure B-1: Armadillo slaying lawyer.

Figure B-2: Armadillo eradicating national debt.

Bibliography

[1] 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.

[2] 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.