Analyzing Linguistic Features of Written Fake and Real News

Alexandra Serralta[a], Ardi Jusufi[b], James Bannon[c], Nishanth Pavinkurve[d]

CSCI-GA.3033-002; Spring 2017

[a] *avs327@nyu.edu*
[b] *aj2223@nyu.edu*
[c] *jjb509@nyu.edu*
[d] *npp267@nyu.edu*

1. Introduction

Accurate decision making requires accurate information. Since the industrial revolution, widely-distributed, written news has been the medium through which a large proportion of the world's population has acquired the information with which they make decisions ranging from the domestic to the political. During the 2016 U.S. Presidential Elections, the topic of "Fake News" became a dominant buzzword for the digital-age version of deception in written news, a phenomenon that has existed at least since 1860s[1], but possibly since the first circulated newspapers.

In our view 'Fake News' is the term for Yellow Journalism in the digital age. Franklin Luther Mott gave five characteristics of Yellow Journalism[2] that show significant overlap with the modern discussion of fake news. These are 1) scare headlines in huge print, 2) lavish use of pictures, 3) use of faked interviews, misleading headlines, and pseudoscience, 4) emphasis on full color Sunday supplements usually with comic strips, and 5) dramatic sympathy with the "underdog" against the system. With these in mind and with reports that fake news was used to aid Donald Trump, [3] and that lower educational levels correlated with counties voting for Trump,[4] we sought to investigate the contrast in quantitative measures of linguistic sophistication between "Fake News" and "Real News."

2. Related Work

Prior research on fake news has focused on two directions: pinpointing the spread and sources of fake news, and analyzing the patterns in its content.

It is commonly agreed that fake news tends to convey a lot of information, claims and emotions in the title, whereas the text body often has only few logically-sound arguments. In certain cases, the body only tangentially agrees with the title. In this context, certain work[5] is centered around labeling whether article titles match their body content, thereby debunking an important problem in fake news. Another research direction tackles the virality of fake news.[6]

Fake news tends to be less linguistically "rich" than real news[7]. This conclusion is made on the basis of stylistic features (such as number of stop words and punctuation), complexity features (sentence syntax tree depth, readability and easiness of comprehension) and psychological features (number of words evoking certain emotions). Overall, it is claimed that fake news target audiences that are unlikely to read much beyond the titles.

3. Hypotheses

3.1. Initial Results and Pivot

Broadly construed, our initial hypothesis was that fake news is less linguistically sophisticated than real news. Mathematically framed: our initial hypothesis was that the distributions of a few linguistic measures have higher means in Real News than they do in Fake News.

3.1.1. Mathematical Formulation

For a quantitative linguistic measure m let $X_f^{(m)}$ be the distribution of fake news in that measure and let $X_r^{(m)}$ be the corresponding distribution for real news. Let $\mathbb{E}\left[\cdot\right]$ be the expectation operator. Then, for all measures m, we have our null hypothesis $\mathcal{H}_0^{(m)}$ and alternate hypothesis $\mathcal{H}_1^{(m)}$ as follows:

$$\begin{aligned} &\mathcal{H}_0^{(m)} := X_f^{(m)} \text{ and } X_r^{(m)} \text{ distributed identically} \\ &\mathcal{H}_1^{(m)} := \mathbb{E}\left[X_r^{(m)}\right] > \mathbb{E}\left[X_f^{(m)}\right] \end{aligned}$$

To test these hypotheses, we used a nonparametric permutation test (as described in Section 4) to determine the significance of our hypothesis.

3.1.2. Initial Results

We tested our hypothesis in the manner described below on the smaller corpus on these metrics: lexical diversity (calculated on both the body and title of the article), Flesch-Kincaid Grade Level (both the body and title of the article), the number of all capital words in the article, and the number of exclamation points in the article body.

In every instance we failed to reject the null hypothesis. We also used a combination of bootstrapping and kernel density estimation to approximate a distribution over p values. The plots showing these are in the Dropbox folder ‘Initial Hypothesis Plots’ for graphical descriptions of our results.

3.1.3. Pivot & Final Results

Realizing that 2048 was only a small portion of the number of possible permutations given the number of articles ($462! \approx 1.4 \times 10^{1032}$; see Section 4), we decided to focus on lexical diversity and sentiment score, which were not too far outside the rejection region and repeat the experiment with $1,000,000$ permutations hoping to capture true convergent behavior. Because articles within sites could plausibly be correlated we chose to also look at the sitewise aggregate SMOG for sites that contributed more than 100 articles to our large corpus. For these three measures we performed a permutation test with 10^6 permutations between 3 and 20 times *each* to make sure our results were stable. In all cases, they were and we rejected the null hypothesis. The plots generated in this way are in the Dropbox folder “Post-Pivot Plots” and summarized in the table below. An example of the plot can be seen in Figure 1.

Measure	**p value**
Mean Lexical Diversity (article body)	≈ 0.042
Mean Sentiment Score (article body)	≈ 0.037
Mean Sitewise SMOG (article body)	≈ 0.018

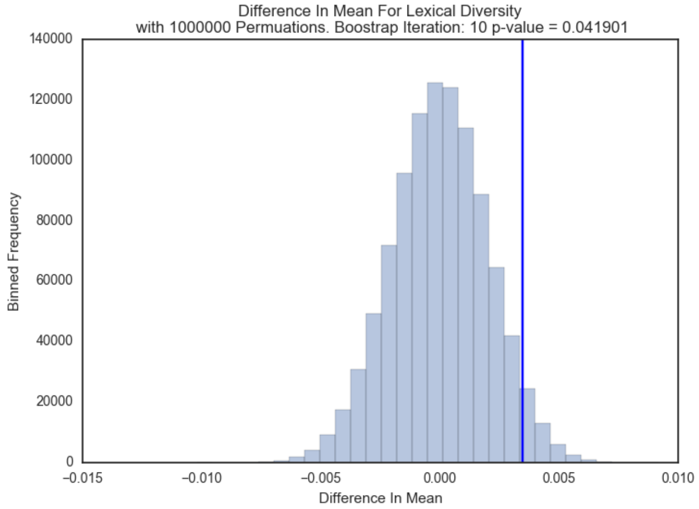

Figure 1: Distribution in Difference of Means of Lexical Diversity for 10th bootstrap iteration. Vertical line is Diff_{lab}

4. Methodology

We gathered examples of real and fake news from several news sources and performed a non-parametric permutation test of our hypotheses following these steps:

1. Fix a significance threshold and a number N of permutations; in our experiment, these were set to $\alpha = 0.05$ and $N = 10^6$, respectively.
2. Compute the labeled difference in means Diff_{lab}.
3. Randomly permute the labels N times and compute $\text{Diff}_1, \ldots, \text{Diff}_N$.
4. Compute the p value as follows:

$$p = \mathbb{P}\left(\text{Diff}_{lab} | \mathcal{H}_0^{(m)}\right) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\text{Diff}_i \geq \text{Diff}_{lab}}$$

   where $\mathbb{1}$ represents the usual indicator function.
5. If $p < \alpha$, reject the null hypothesis; otherwise, we fail to reject[8].

4.1. Gathering the list of news sources

In order to obtain a list of real and fake news sources, we referred to an independent fact-checking website, MediaBiasFactCheck.com(MBFC). The initial dataset scraped from the website had the following features: **Name**: The name of the news website, **Link**: The link to the main page of the website, **Bias**: The label assigned to the news source by MBFC according to their methodology[9]. The label assigned takes values in: {*Least Biased*, *Right-Center*, *Left-Center*, *Right*, *Left*, *Extreme Left*, *Extreme Right*, *Questionable Sources*, *Conspiracy Theory/Pseudoscience*}, and **Factual Reporting Score**: The score determined by MBFC as described in their methodology. The score assigned takes values in: {*Very High*, *High*, *Mixed*, *Low*}. Following Mott's third criterion, sites listed as "Questionable Sources" or "Conspiracy Theory/Pseudoscience" were given the label "fake", while all others were called "real."

4.2. Article extraction

We developed a Python script for extracting articles from each news source. The script was designed such that from the main page, all links potentially containing articles were scanned through, and from each successful article page, related links were considered in a Breadth-First Search style approach, until up to 200 articles were collected.

We generated two corpora, a small one consisting of 462 articles from 54 websites, with 1 to 10 articles extracted from each, and a large one consisting of 149,572 articles from 1,104 websites, with 1 to 200 articles from each. The following data was stored for each article in both corpora: *Mainpage URL, Article URL, Bias, Factual Reporting Score, Article Title, Article Text, Authors,* and *Publishing Date*.

4.3. Metrics

Once the database of articles was constructed, a set of metrics was used in order to analyze lexical qualities of fake and real news. These metrics were used on both the title and text of an article separately.

4.3.1. Lexical Diversity

Lexical Diversity is defined as the number of unique words in a text to the total number of words. This was calculated using the NLTK word tokenization functionality.

4.3.2. SMOG

The SMOG grade is a measure of readability that estimates the years of education needed to understand a piece of writing. SMOG is an acronym for Simple Measure of Gobbledygook. The equation used is as follows:

$$\text{grade} = 1.0430\sqrt{\text{number of polysyllables} \times \frac{30}{\text{number of sentences}}} + 3.1291 \tag{1}$$

4.3.3. Flesch Kincaid Grade Level

The Flesch Kincaid tests quantify the difficulty of comprehension of a passage in English by assigning it a score, which then is translated to an educational level from 5th grade to College Graduate level. Applying this measure to real and fake news would allow us to see possible differences in complexity within the two types of news. The equation used to attribute these scores is as follows:

$$0.39\left(\frac{\text{total words}}{\text{total sentences}}\right) + 11.8\left(\frac{\text{total syllables}}{\text{total words}}\right) - 15.59 \tag{2}$$

4.4. Sentiment Analysis

Sentiment Analysis is a branch of Natural Language Processing that seeks to assign an overall sentiment to a text. Most Sentiment Analysis systems separate text into three different categories: positive, negative and neutral. This metric was chosen to analyze the emotional character of Fake News versus Real News. Once again, the NLTK library was used with its Sentiment Analyzing system called Vader.[10]

5. Discussion & Potential For Future Work

Our analysis is fundamentally limited in scope. We analyzed only English-language articles and our hypotheses were generated based on linguistic features that we suspected due to the U.S. 2016 presidential election. Since the motivations for generating fake news likely vary across websites, and perhaps even countries and cultures, future work could involve other metrics and/or different data sets and, perhaps more interestingly, could tap into distributional networks that spread fake news to try and discover different groups spreading different kinds of deception for disparate reasons.

[1] Techniques of 19th-century fake news reporter teach us why we fall for it today:
`https://theconversation.com/techniques-of-19th-century-fake-news-reporter-teach-us-why-we-fall-for-it-today-75583`

[2] Yellow Journalism:
`https://en.wikipedia.org/wiki/Yellow_journalism`

[3] Russia hired 1,000 people to create anti-Clinton 'fake news' in key US states during election:
`http://www.independent.co.uk/news/world/americas/us-politics/russian-trolls-hilary-clinton-fake-news-election-democrat-mark-warner-intelligence-committee-a7657641.html`

[4] Education, Not Income, Predicted Who Would Vote For Trump:
`http://fivethirtyeight.com/features/education-not-income-predicted-who-would-vote-for-trump/`

[5] Fake News Challenge:
`http://www.fakenewschallenge.org`

[6] Hoaxy: A Platform for Tracking Online Misinformation:
`http://dl.acm.org/citation.cfm?doid=2872518.2890098`

[7] This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News:
`https://arxiv.org/pdf/1703.09398.pdf`

[8] Hypothesis testing. Non Parametric Testing: the Permutation Test:
`http://math.nyu.edu/ cfgranda/pages/DSGA1002_fall16/material/hypothesis_testing.pdf`

[9] Media Bias Fact Check - Methodology:
`https://mediabiasfactcheck.com/methodology/`

[10] VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text:
`http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf`