**Never Judge an Article by its Title:**

**An Exploratory Analysis of Headlines in Real and Fakes News Articles**

Navraj Narula

Department of Journalism

Columbia University – New York, NY USA

nnn2112@columbia.edu / navrajnarula@gmail.com

**Abstract**

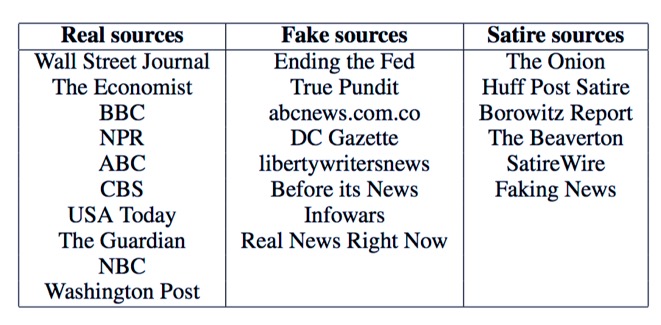
Jim Morrison, an American songwriter, once said: “Whoever controls the media, controls the mind.” Media—in its many televised, broadcasted, and printed forms—not only acts a major influencer in our everyday lives, but is often labeled as a harbinger of truth for many. In the stunning election results that stunned the world in late 2016, attention in regards to fake news has increased. Often times, signs of ingenuity are even present in the headlines of news articles themselves. I examine \_\_\_\_\_ known real and fake to determine linguistic differences present in them both. My results indicate that \_\_\_\_

**Introduction and Paper Setup**

**Dataset and Usage**

The dataset that I am using to inform my classifier was made available to me by Benjamin D. Horne and Sibel Adali, a PhD student at Rensselaer Polytechnic Institute’s Social Cognitive Network Academic Research Center (SCNARC) and a professor of computer science at the aforementioned institution. The dataset itself is self-collected and contains news articles most pertinent to past election cycle, in which Donald Trump won the 45th United States presidency with 306 electoral votes [1].

In total, they obtained 75 randomly-selected news articles each with associated headlines for three categories: real news, fake news, and satirical news. Horne and Adali used Zimdars’ list of fake news sources and Business Insider’s list of “most trusted” news sources to construct their datasets [2,3]. In regards to satirical sources, they used websites that openly stated that they were a satirical news source on the front page. Below is a table of categorized news sources retrieved from their paper [4]:



For my analysis, I have chosen to focus only on data related to real and fake sources. In particular, I am zooming in on news article headlines rather than story content. As reported by Chris Cilliza and according to the Media Insight Project study conducted by the AP-NORC Center for Public Affairs Research and the American Press Institute, six in 10 people admit that they do not read beyond a news article headline and “in truth, that number is almost certainly higher than that, since plenty of people won’t admit to just being headline gazers, but, in fact, are” [5].

Little time dedicated to verification of simply a headline can prove harmful, especially when social sharing is a factor at play. As a final result, I was properly able to obtain 63 real news headlines and 72 fake news headlines from Horne and Adali’s manually-constructed dataset of randomized political articles.

I reorganized all obtained files into a table with two columns: one for the text of the article headline, and one for the category it pertains to (i.e. a binary label of “real” or “fake”).

Here is example of a real and fake news article, as seen in the dataset:

|  |  |
| --- | --- |
| Text | Category |
| Walmart pulls 'Black Lives Matter' shirts from website after cop complaints | real |
| Obamas Racist Attacks Against White Working Class Caused Historic Democrat Party Collapse | fake |

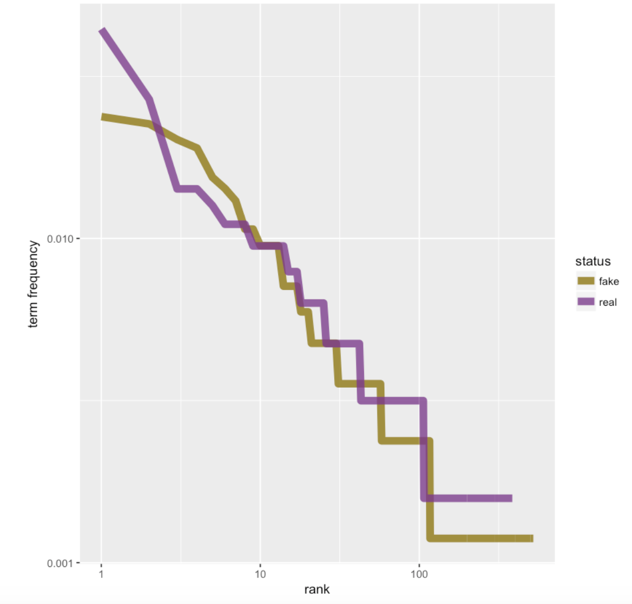
**Experiments and Results**

There exist multiple methods to derive meaning from text. In my paper, I have conducted an exploratory analysis of three methods—one visual, one algorithmic, and one more so human.

\*

In order to understand the content present in the headlines at a naïve level, I first calculated the term frequency for each word present in the headline. Given the fact that the dataset revolved around a very niche topic, it was not surprising to see that the top words for both real and fake news headlines were the names of politicians. In real news headlines, the word “trump” appeared 28 times out of 633 words. In fake news headlines, the word “obama" appeared 20 out of 842 times. Following this, the word “trump” appeared 19 times.

After obtaining the frequency of each word, I utilized the concept of Zipf’s law to model the distribution of terms by plotting the rank of each word against its frequency [6].



From the above graph, we can see that the slope is negative as the frequency decreases and the rank increases for each word. Because the number of real and fake news headlines is not equal in our dataset, the last rank for real news headlines halts before the rank for fake news headlines. It is interesting to note, though, that the rank for real news headlines at the start of the graph is higher than that for fake news headlines—again, with the top ranking word for real news headlines as “trump” and the top ranking word for fake news articles as “obama.”

In order to get a more accurate picture regarding the content of real and fake news headlines for my dataset, I applied TF-IDF (term frequency-inverse document frequency) to obtain the most relevant for the text present in each headline instance. According to Silge and and Robinson, authors of *Text Mining with R*, TF-IDF is accomplished by “decreasing the weight for commonly used words and increasing the weight for words that are not used very much in a collection or corpus of documents” [7].

In our case, TF-IDF will attempt to find words that are common in our corpus of headlines; however, “not *too* common” [7].

\*

**Evaluation**

**References**

[1] “Presidential Results.” *CNN Politics*. Retrieved: Dec-15-2017. URL: <http://www.cnn.com/election/results/president>

[2] Zimdars, Melissa. “False, Misleading, Clickbait-y, and/or Satirical “News” Sources.” Retrieved: Dec-15-2017. URL: <https://docs.google.com/document/d/10eA5-mCZLSS4MQY5QGb5ewC3VAL6pLkT53V_81ZyitM/preview>

[3] Engel, Pamela. “Here Are The Most- And Least-Trusted News Outlets In America.” *Business Insider.* Retrieved: Dec-15-2017. URL: <http://www.businessinsider.com/here-are-the-most-and-least-trusted-news-outlets-in-america-2014-10>

[4] Horne, Benjamin D., Adali, Sibel. “This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News.” *Rensselaer Polytechnic Institute*. The 2nd International Workshop on News and Public Opinion at ICWSM. Retrieved: Dec-14-2017.

URL: <https://arxiv.org/pdf/1703.09398.pdf>

[5] Cilliza, Chris. “Americans read headlines. And not much else.” *The Washington Post*. Retrieved: Dec-15-2017.

URL: <https://www.washingtonpost.com/news/the-fix/wp/2014/03/19/americans-read-headlines-and-not-much-else/?utm_term=.155c7f12290b>

[6] “Zipf's law: Modeling the distribution of terms.” *Stanford University Natural Language Processing Group*. Retrieved: Dec-15-2017.

URL:

<https://nlp.stanford.edu/IR-book/html/htmledition/zipfs-law-modeling-the-distribution-of-terms-1.html>

[7] Silge, Julia., Robinson, David. “Analyzing word and document frequency: tf-idf.” *Text Mining with R: A Tidy Approach*. Retrieved: Dec-15-2017.

URL: <https://www.tidytextmining.com/tfidf.html>