Never Judge an Article by its Title:
An Exploratory Analysis of Headlines for Real and Fakes News Articles

Navraj Narula
Department of Journalism
Columbia University – New York, NY USA
nnn2112@columbia.edu / navrajnarula@gmail.com

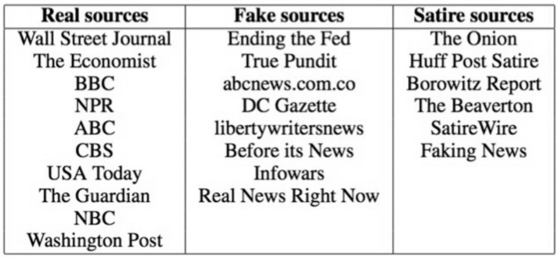
Introduction

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 as 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. In my paper, I examine 135 labeled real and fake news headlines in the context of visualization, classification, and general interview conduction. My results indicate that distinction of short texts (i.e. headlines) is challenging to categorize for both machines and humans; however, linguistic differences appear visible to both parties.

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 the past election cycle, in which Donald Trump won the 45th United States presidency with 306 electoral votes [1].

In total, Horne and Adali 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]:



Real sources	Fake sources	Satire sources
Wall Street Journal	Ending the Fed	The Onion
The Economist	True Pundit	Huff Post Satire
BBC	abcnews.com.co	Borowitz Report
NPR	DC Gazette	The Beaverton
ABC	libertywritersnews	SatireWire
CBS	Before its News	Faking News
USA Today	Infowars	
The Guardian	Real News Right Now	
NBC		
Washington Post		

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. In total, I have 135 news headlines.

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 reconstructed 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

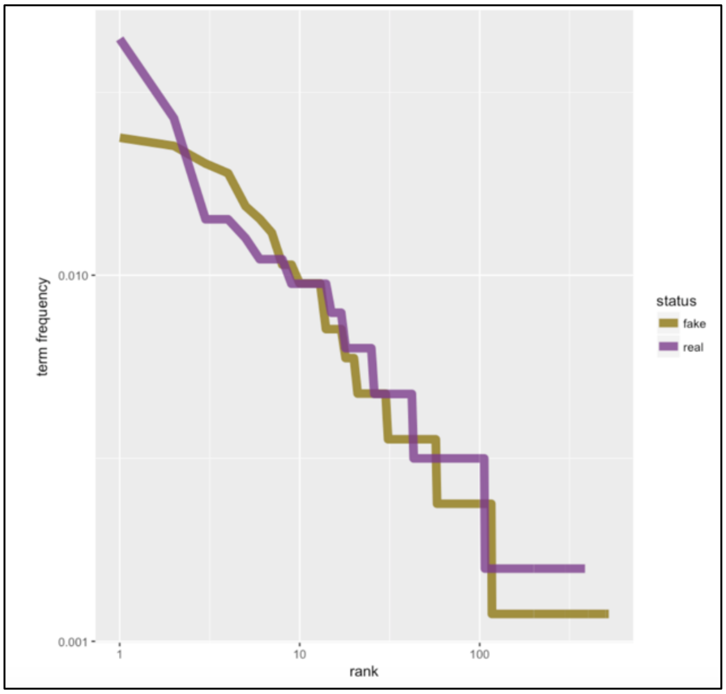
Experiments and Results

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 in fake news headlines.

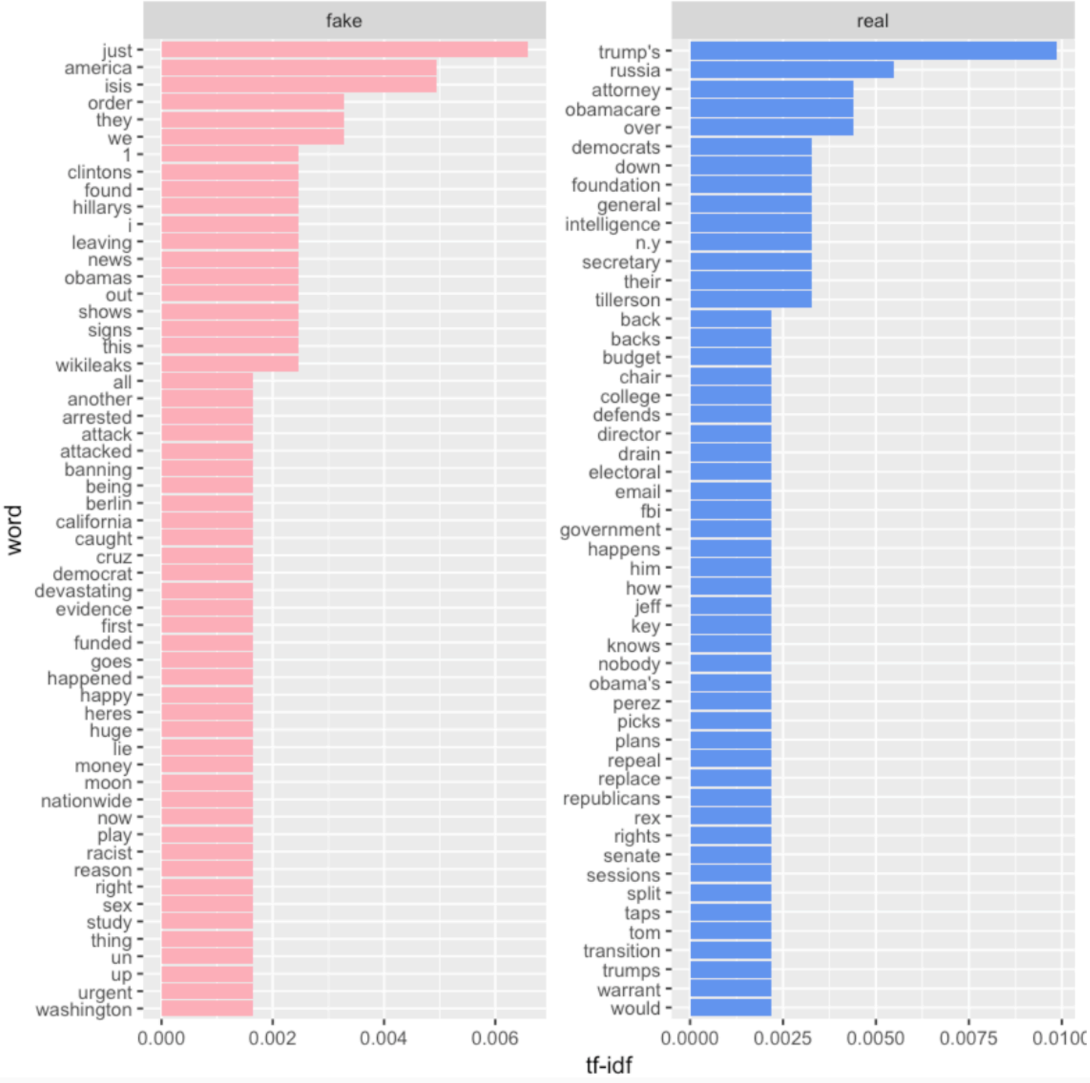
After obtaining the frequency of each word, I utilized the concept of Zipf’s law to model the distribution of terms by plotting the rank of each word against its frequency [6].



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 the 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 headlines 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 words 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]. Below I have included visualizations for the 25 most common words, according to the TF-IDF algorithm:



In comparing the two, we can see that words in fake news headlines are more so simplistic than words in real news headlines.

Some of which include “just,” “they,” and “i.” Further down the chart, we have “happened” and “thing.” Several fake news headlines that include such words are: “BREAKING: We just found out who attacked TRUMP,” “Ted Cruz: I Will Endorse Donald Trump For President If He Makes Masturbation Illegal,” and “WIKILEAKS DOCUMENTS REVEAL APOLLO PROGRAM WAS A FRAUD, MOON LANDINGS NEVER HAPPENED.”

Words in real news headlines include “attorney” and “over.” Further down, we have “rights,” “transition,” and “would.” Real news headlines in the dataset that include these words are: “Trump Plans To Dissolve His Foundation; N.Y. Attorney General Pushes Back,” “Trump again claims he 'would have won' popular vote',” and “GOP in Congress split over how to repeal Obamacare.”

Words in fake news headlines contain language that tend more so to the extreme as well as include absolute statements. Real news headlines have a softer connotation and tend to not reveal an entire story in the headline, encouraging the reader to read more about the facts.

Considering words in isolation reveals a lot of information; however, at times, valuable information in regards to capitalization, punctuation, and modification can be lost. A survey analysis further in the paper will take such information into consideration; however, this naïve tactic in examining words in isolation do provide preliminary insights in regards to headline content and design for real and fake news articles.

*

Journalism is work done with the interest of the public in mind. In order to avoid the spread of misinformation, newsrooms and other organizations that seek to further the mission of truth may utilize classification algorithms that fall under the umbrella of machine learning to categorize large quantities of documents.

In regards to preprocessing the headlines, I utilized R’s text mining library to:

- Lowercase each word
- Remove all numbers
- Remove all English stopwords
- Remove punctuation
- Strip off whitespaces
- Stem all words

I then divided my headlines into a train and test set, both of equal value. After doing so, I used the K-Nearest Neighbors algorithm to make a prediction on the test set regarding whether or not a given headline was real or fake. After running this algorithm a thousand times exactly, the average accuracy returned was 64.5%. The highest accuracy achieved during these rounds, however, was 79.1%. The confusion matrix for this instance is provided below:

	Actual	
Predictions	fake	real
fake	24	4
real	10	29

Out of 67 headlines from the test set, 80% of fake news headlines were classified correctly while 74.3% of real news headlines fell into the right category. The slight dip in percentage for real news headlines may have been due to the fact that lesser real news headlines made their way into the training set. Nonetheless, despite the small sample of headlines available in the dataset, the K-Nearest Neighbors algorithm performs at best average given a thousand runs of the model yielding an accuracy just shy of 65%.

*

In addition to experimenting with a computational classifier, I asked people themselves to classify whether or not they thought a given headline was real of fake. I arbitrarily extract five real news headlines and five fake news headlines from the dataset and asked people to indicate whether or not the headline was real or fake. I received a total of 121 responses and have included my breakdown of the results in the table below:

Headline	Actual Label	Correct	Incorrect
Recent Study Shows Nearly Six in Ten Trump Supporters are Illiterate	Fake	86%	14%
REPORT: Nine ISIS Supporters Arrested Near Washington D.C.	Fake	63.6%	36.4%
Will Trump's plan mean faster rate hikes?	Real	76%	24%

Ahead of Pro-Trump Rally, KKK Members Claim They're 'Not White Supremacists'	Real	67.8%	32.2%
Obamacare Has Its Biggest Day as Republicans Promise Repeal	Real	71.9%	28.1%
Obama To Issue Executive Order Extending Presidential Term Limits	Fake	94.2%	5.8%
Trump Pledges to End "Let's Move" on First Day in Office	Fake	67.8%	32.2%
Carl Paladino, Trump campaign N.Y. co-chair, says his derogatory Obama remarks were a "mistake"	Real	72.7%	27.3%
Obama "permanently" bans drilling in parts of Arctic, Atlantic	Real	39.7%	60.3%
Mysterious Manhattan Military Flyover Was Trump Rescue Exercise	Fake	78.5%	21.5%

People who engaged in my survey were able to correctly distinguish fake news headlines from real news headlines 90% of the time, exceeding the accuracy of my classifier by about 25%. The only headline from my survey that misled people was: “Obama "permanently" bans drilling in parts of Arctic, Atlantic.” This was actually a real news headline, but 60.3% people indicated that it was fake. People were most able to correctly classify “Obama To Issue Executive Order Extending Presidential Term Limits” as a fake headline. As for real headlines, the title that people were most able to correctly identify was “Will Trump's plan mean faster rate hikes?” [8].

In addition to asking people to classify headlines, I also asked them to describe whether or not distinguishing fake news headlines from real news headlines was difficult. Many people reported that the task was, in fact, difficult because “the language seemed pretty similar in both [headlines]” or that “the language used in each heading did not seem all that different to me.” One person mentioned that they may have had a tough time telling the difference because they pay little attention to the news [8]. Some people did not find the task difficult at all seeing as they generally do keep up with the news or do not always assume that headlines contain the entire truth—and therefore, read on.

In comparison to real news headlines, people state that fake news headlines contain language that tends to incite an emotional response. It could be “alarmist…[or] extremes expressed through definitives, sensationalism,” or rather “exaggerated language that clearly isn’t self-aware of its exaggeration.” One person stated that if a headline is more “well-written” for a fake news article, then the statement in the headline may refer to “blatantly false event” such as with the headline “Obama To Issue Executive Order Extending Presidential Term Limits,” in which 94.2% of people indicated as fake due to the absurdity associated with the statement [8].

Despite minor error, it is no surprise that humans are able to make such decisions intelligently in comparison to machines.

Reflection

In terms of working with textual data, I found it helpful to explore and experiment with the different types of methods discussed in this paper. For a more focused piece, I might want to perform an extensive user-research type study in which I ask questions not only related to headline categorization, but also about the demographic of the reader. I may also want to expand my dataset to include more content so as to gain better insight in regards to whether or not linguistic differences affect a certain audience over another. This could be used, in turn, to tune my classifier for more telling results and an increased accuracy.

Different results will arise from using different methods—or even a different dataset. Finding truth in that matter is also a factor to consider when performing such analyses.

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

[8] Narula, Navraj. "Real of Fake News Headlines?" *Google Survey*. Created: Nove-23-2017. Retrieved: Dec-15-2017. URL:
https://docs.google.com/forms/d/1Qu47uVbD4ol8w_RsjeX1EOT0Ro0uqKvHTHvMImdtlvY/edit