Team False Positives: Omar Khan, Daniel Snitkovskiy, Dominick Tavitian, Nattanon Bunyatipanon
INFO 370: Data Science
Lavi Aulck & Li Zeng

TEAM FALSE POSITIVES: FINAL PAPER

Team False Positives: Omar Khan, Daniel Snitkovskiy, Dominick Tavitian, Nattanon Bunyatipanon
INFO 370: Data Science
Lavi Aulck & Li Zeng

Table of Contents

Team False Positives: Omar Khan, Daniel Snitkovskiy, Dominick Tavitian, Nattanon Bunyatipanon
INFO 370: Data Science
Lavi Aulck & Li Zeng

Previous Work

Although we were unable to find any scholarly articles pertaining to studies of fake news, there have been technologies that have been released to aid in its identification and classification. One of the works that we found was a browser extension called 'bs detector'. It is built upon an open source platform that checks the visited URL to a manually curated list of domains that have been reported or have a history of spreading fake news. Although there is a disclaimer on their homepage that they do not retain ownership or responsibility of the list of fake domains any longer, we would still be able to utilize that list as one of the possible point of data to analyze news articles. (B.S. Detector 2017)

Another prominent fake news tool is Reuter's 'Tracer News'. It does the opposite of what we are trying to accomplish in a sense that instead of detecting fake news, it detects real news on Twitter. According to internal research from Reuters, about 20 percent of news break on Twitter before mainstream media is able to get ahold of it. It was also recently discovered in the latest election cycle that fake news can spread as fast as real news. This tool's algorithm is proprietary and has over 700 signals to determine whether trending topics are truthful. Although there isn't data on its false positive rates, it does put into perspective the gravity of how influential fake news can be in determining political outcomes. (Iozzio 2016)

Motivation

According to Pew Research, the number of average monthly unique users for the "highest-traffic digital-native news outlets" has grown at an rate of 1,913,216/year between 2014 and 2016 (Stocking 2017). Unfortunately, with the rise of online news consumption, the bar for accountability and validity has been replaced with a focus on clicks and shares. According to a 2016 report by Stanford researchers, fake news has grown because of the "[lowered] barriers to entry...social media are well-suited for fake news dissemination... a decline of 'trust and confidence' in the mass media…[and] the rise of political polarization…" (Allcott et al 2017) The volume of fake news and the ease of dissemination make Americans who do not utilize critical thinking and statistical skepticism (i.e. 'calling BS') susceptible to believing falsehoods. This susceptibility is the reason why we want to better understand fake news, and, ultimately, detect it. By providing a mechanism to detect unreliable news content, falsehoods will be less likely to spread, and a greater sense of trust in the credibility and validity of news outlets can begin to be rebuilt.

Team False Positives: Omar Khan, Daniel Snitkovskiy, Dominick Tavitian, Nattanon Bunyatipanon
INFO 370: Data Science
Lavi Aulck & Li Zeng

Research Question

Our exact research question is “With the content of news articles, can we use NLP techniques to correctly classify if the news reported in an article is fake or not?” NLP (Natural Language Processing) deals with processing language and text data to solve a variety of problems like finding duplicate questions in forums like Quora and Stackoverflow, or translating between different languages.

Data

We began by scraping 1200 articles across the news sources of the Economist, PBS, NPR and BBC. After conducting our preliminary analysis, we realized that our fake news data was heavily skewed towards political articles, while the “real news” articles that we scraped varied wildly in terms of categories. The implication of this was that any models that would be built using these data would be biased (i.e. applying the “real” label to articles because they are non-political, and vice-versa). Therefore, for this analysis, we decided to rescrape about 600 articles, focusing only on articles that are categorized as political. Another source of bias was the language of the fake news dataset. Since all of the scraped articles were in English, and there were non-English articles present only in the fake news dataset, we removed the non-English rows from said dataset to prevent language from being a lurking variable in our model. With these modifications, the focus of our learner will be primarily on the textual content of the articles.

Team False Positives: Omar Khan, Daniel Snitkovskiy, Dominick Tavitian, Nattanon Bunyatipanon
INFO 370: Data Science
Lavi Aulck & Li Zeng

Analysis

Data Cleaning

Data cleaning included the process of transforming the raw text of an article (i.e. the "text" attribute) into a shorter and more information rich format. First, the text was normalized by converting it to lowercase and filtering out non-ASCII characters. Second, the "stopwords" (words that serve grammatical purposes but lack information) were taken out. Third, the words were converted to their linguistic roots through a process called "stemming" (e.g. the root of the words "loved" and "loving" is "love) in order to reduce redundancy. Finally, to protect against erroneous results, only words part of the "wordnet" corpus from nltk were kept in the final result.

Featurization

There are two main strategies employed to convert the text of each article into information rich inputs for our classifier: Bag-of-Ngrams and TF-IDF.

Bag-of-Ngrams modeling is a framework for textual representation that creates a term frequency matrix based on a specified token (i.e. an "N-gram" or set of consecutive words). That is, for a specified term (e.g. unigram or "singular word"), the Bag-of-Ngrams model would be a two dimensional matrix with the observations (in this case, articles) as the rows, and each unique term as the columns. For row i and column j, the (i,j)'th entry represents the number of occurrences of token j in observation i.

For our analysis, we are utilizing two variants of the Bag-of-Ngrams: Bag-of-Words (i.e. unigrams) and Bag-of-Bigrams (i.e. word pairs). Bag-of-Words was chosen because our classifier is linear (i.e. logistic), and empirical analysis has shown that "...bag-of-words model[s]...capture sufficient information for linear classifiers to make highly accurate predictions." (Heap et. al, 2017). Since the tokens are individual words, one also retains a high degree of interpretability by being able to make statements on what words are particularly "influential". Bag-of-Words, however, does have its limitations. The main limitation is that the context and word ordering is lost in the process of extracting the unigrams from entire articles. To retain more of this structure, we utilize a "Bag-of-Bigrams" to maintain words that are directly adjacent, so that the idea of "phrases" can be represented directly. This does with two distinct costs: interpretability and computational costs. Although bigrams retain more of the sequential structure of the origin, these bigrams may not always be semantically meaningful phrases (i.e. collocations). In the worst case, for "k unique words, there could be k^2 unique...bigrams..." (Zheng, 2017).

In order to reduce the sparseness of both of these representations (and therefore, training time), only the top 1000 most frequent tokens were retained as columns in the final matrices. Scikit-learn's "CountVectorizer" function was used in order to facilitate these conversions.

Although Bag-of-Ngrams retains sufficient information, it utilizes the raw counts to represent term frequency, which can incur a) memory costs and b) give undue significance to commonly occurring, less information rich terms. In order to prioritize rarer, semantically meaningful tokens, the Bag-of-Words and Bag-of-Bigrams representations were scaled using a Term-Frequency - Inverse-Document-Frequency transformation. According to a 2015 survey of recommender systems, TF-IDF was "the most frequently applied weighting scheme" for "content-based" filtering. (Beel et. al, 2015). The drawback of TF-IDF is the added computational cost of scaling the existing term frequencies.

"Term Frequency" is analogous to the entries of the Bag-of-Ngrams matrix, while "Inverse-Document-Frequency" is the inverse proportion of occurrences of the term scaled by the number of observations. That is, for token t and number of articles n, IDF(t) = (number of articles with t / n). In order to scale down this term and make frequent terms tend to 0, the logarithm of the IDF is taken. A "+1" term is also appended to the IDF to prevent division by 0. Lastly, instead of using the raw counts for the term frequencies, the l2 norm of the counts are taken in order to make the model "quicker to train..."(Zheng, 2017) Scikit-learn's "TfidfTransformer" function was used in order to facilitate these conversions.

In summary, for our classifier, we explored four distinct feature representations: Bag-of-Words, Bag-of-Bigrams, TF-IDF transform for Bag-of-Words, and TF-IDF transform for Bag-of-Bigrams.

Team False Positives: Omar Khan, Daniel Snitkovskiy, Dominick Tavitian, Nattanon Bunyatipanon
INFO 370: Data Science
Lavi Aulck & Li Zeng

Summary of Data

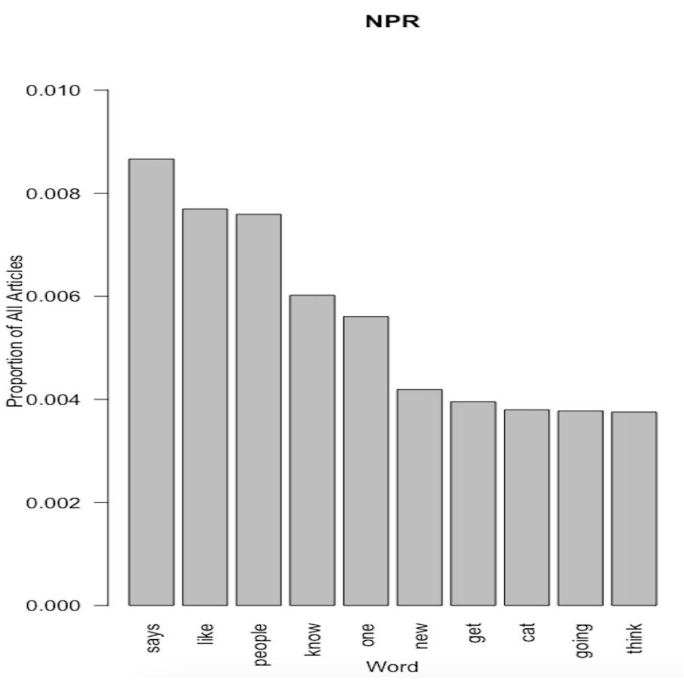
The following analysis concerns the Fake News dataset:

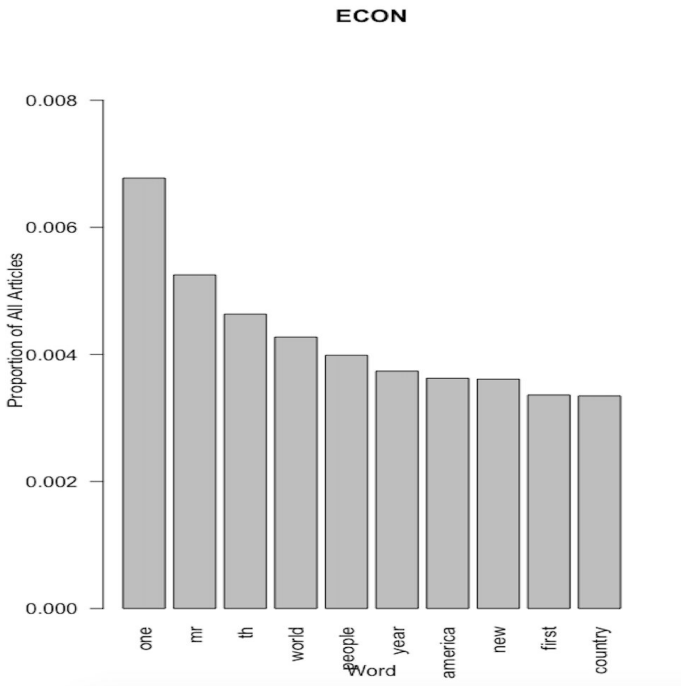
Column	Description	Measure of Centrality	Proportion of Observations
language	The primary language of the article	English (mode)	95.4%
country	Country of origin of the article's publication	United States (most frequent) United Kingdom (second most frequent)	79.8% 6.4%
spam_score	Measure of "spamminess" as queried from bs detector	0 (median)	N/A
type	Subcategory of fake news articles	"Bs" (i.e. completely fabricated) (mode)	88.0%

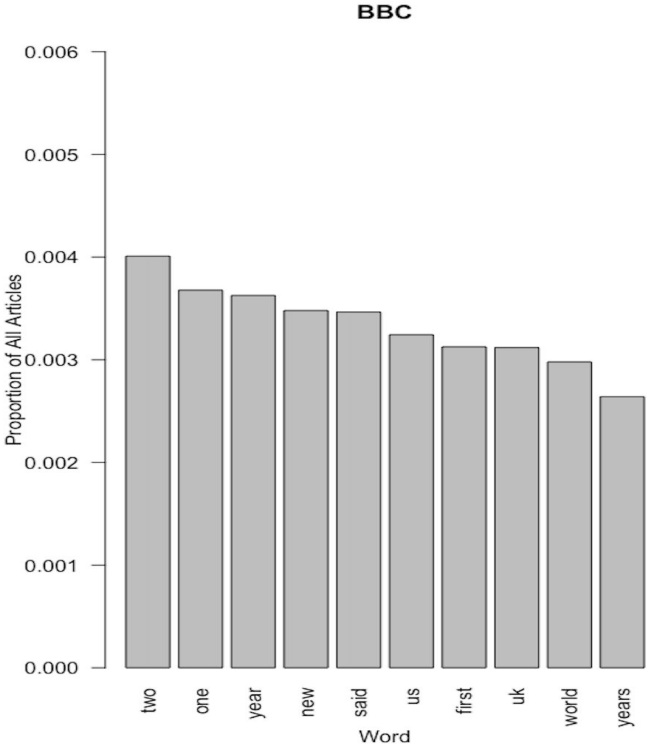
What this shows is that most of the fake news being analyzed is wholesale fabrication (i.e. with the intent to prank), with predominantly English speaking publishers. The proportion of English articles coming from English speaking countries will inform both the cleaning and collection of data in order to ensure language or country of origin do not have unintended impacts on the model. At this time, spam_score and type will not be incorporated, and we will utilize random subsampling during modeling to account for these variations.

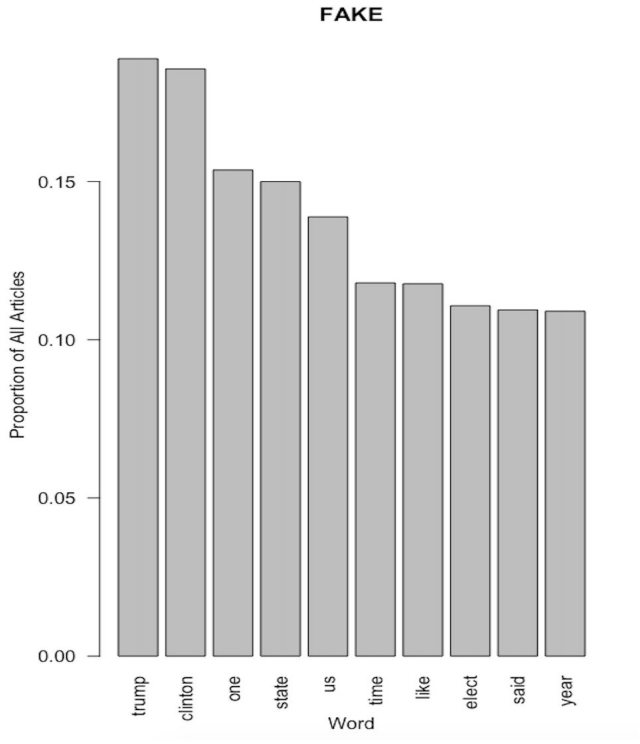
The following series of plots and table depict the distribution of words across all data sources from our initial scrape.

Team False Positives: Omar Khan, Daniel Snitkovskiy, Dominick Tavitian, Nattanon Bunyatipanon
INFO 370: Data Science
Lavi Aulck & Li Zeng







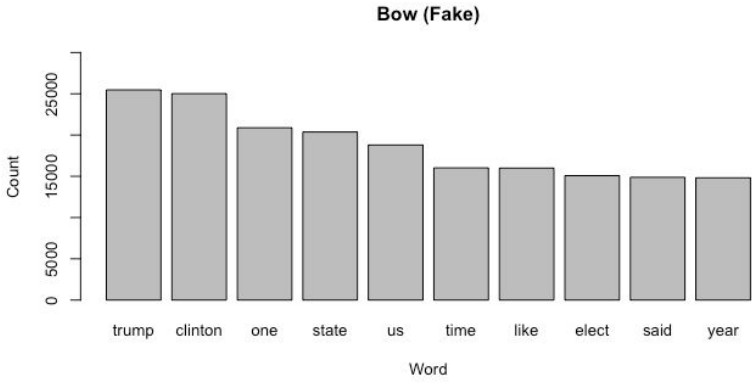


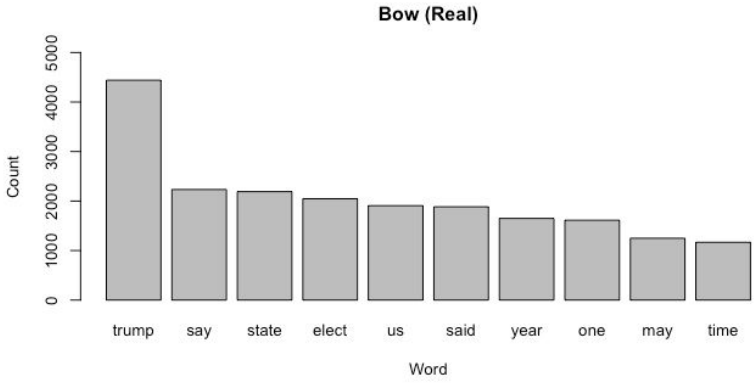
Team False Positives: Omar Khan, Daniel Snitkovskiy, Dominick Tavitian, Nattanon Bunyatipanon
INFO 370: Data Science
Lavi Aulck & Li Zeng

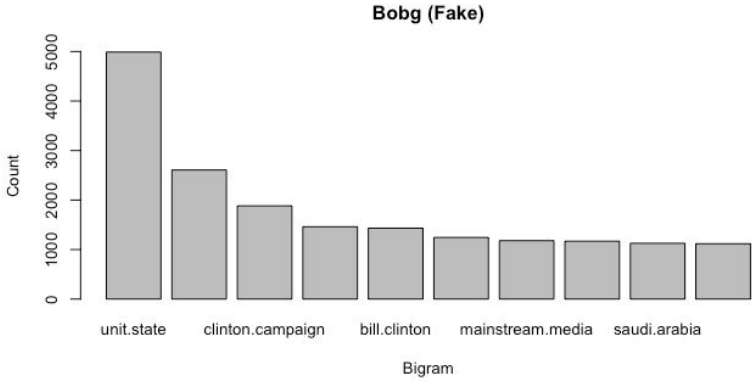
Ranking by Proportion	**Source**			
	NPR	**BBC**	**The Economist**	**Fake**
1	Says	Two	One	Trump
2	Like	One	Mr	Clinton
3	People	Year	Th	One
4	Know	New	World	State
5	One	Said	People	Us
6	New	Us	Year	Time
7	Get	First	America	Like
8	Cat	Uk	New	Elect
9	Going	World	First	Said
10	Think	Years	Country	Year

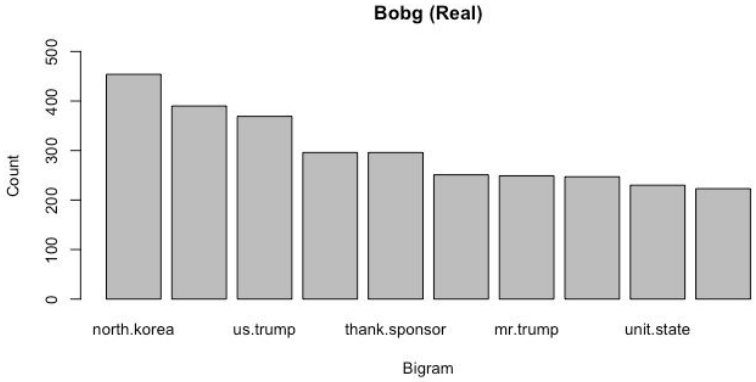
As mentioned in our data section, after conducting our this analysis we realized that our data was heavily skewed towards political articles, while the “real news” articles that we scraped varied wildly in terms of categories. This was remedied by scraping from the politics section of our real news sources. The distributions of this new data scraped is depicted below for each data representation.

Team False Positives: Omar Khan, Daniel Snitkovskiy, Dominick Tavitian, Nattanon Bunyatipanon
INFO 370: Data Science
Lavi Aulck & Li Zeng









Modeling and Results

The following section evaluates the performance of our four representations against the baseline model (i.e. random guessing based on proportion).

For the purpose of modelling we first took our entire set of data and divided into two sets: 1) train + val and 2) test. Within the first set, we took 180 data instances of each class to create a 360 instance pure training set, the remaining data in this set was used for validation. For the purpose of this analysis we only used the train + val set and obtained all our results from the validation subset, our test set remains untouched.

To establish a dummy baseline we used the proportion of each class seen in the training set as our random guessing weights, in this case because our training set had an equal number of occurrences between classes the random guessing was 50:50. This resulted in a model that was approximately 50% accurate, which is to be expected in a binary classification setting for a dummy model. We then trained 4 different logistic regression models using Scikit-learn, each one using a different representation of our data (see below figure for visualized results). All models used l2 loss, and the L-BFGS optimizer, a pairing that has been shown to be fairly efficient for training (Minka, 2003). The highest weighted bigram is

"http www", looking through our dataset we've found many instances of urls in the textual contents of fake news articles. The highest weighted unigram is "video", the highest absolute weighted unigram is negatively weighted and is "sponsor".

Our results, while seemingly good come with an extremely important caveat: even though our training set was completely balanced our validation set is skewed towards fake news, about 90% fake news, so while we observe 90+% F-scores across most models our average misclassification rate (real news classified as fake news) is around 20% while the rate at which we fail to classify (fake news classified as real news) is about 2%, both of these numbers are much higher for Bigram tf-idf model, which performed worse than random guessing.

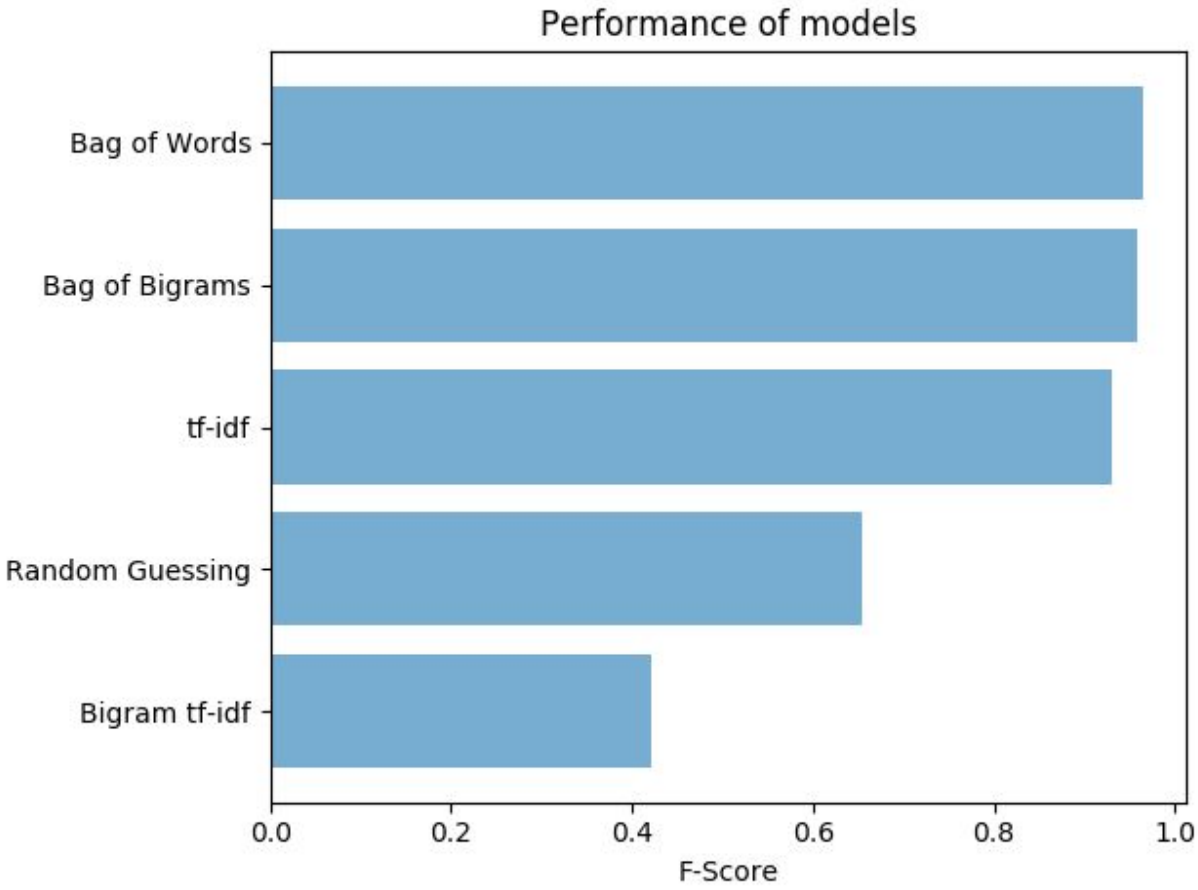


Figure: The performance of the logistic regression on different representations + random guessing.

In summary these modeling results are promising and indicate a very tractable problem, but there is a strong need for a much larger amount of data. Our next steps include acquiring this data, re evaluating our models, interpreting our new models, and trying more complex models if necessary.

Team False Positives: Omar Khan, Daniel Snitkovskiy, Dominick Tavitian, Nattanon Bunyatipanon
INFO 370: Data Science
Lavi Aulck & Li Zeng

Reflection

The class imbalance of our validation sets were briefly mentioned in the analysis section, so in order to validate that these results hold for a more balanced validation set we constructed validation sets that had 50:50 balance between the classes by sampling from the set of data we held out (all non-training data). For 5 different validation sets of this nature for each data representation we computed the mean F-scores and found them to be slightly lower than the validation set that was skewed towards fake news, but the overall trend shown in the above figure held. This tells us that the difference in score across representations was not caused solely by class imbalance.

Distribution of Labor

Task	Members
Set-up Github Repo	Nattanon
Scrape Articles	Nattanon, Dominick
Clean Data	Nattanon, Dominick
Exploratory Analysis	All members
NLP Models, Classification	Omar, Daniel
Analysis with Algorithms	Omar, Daniel
Data Visualization	All members
Finalize Project	All members
Present	All Members

Team False Positives: Omar Khan, Daniel Snitkovskiy, Dominick Tavitian, Nattanon Bunyatipanon
INFO 370: Data Science
Lavi Aulck & Li Zeng

Future Directions

In order to make this work better in the future, we believe if there was a larger volume of high quality data then we could better tackle this problem. This larger volume of high quality data along with more computational resources would enable us to use a larger variety of models like random forests and deep learning, which would help us attain higher accuracy. Though we were able to test some of these models we didn't have enough time/resources to tune these models to higher degrees of accuracy than our logistic regression based models, but with all that's going on in the world, we expect curious data scientists to use these more time-intensive models to solve this problem.

Team False Positives: Omar Khan, Daniel Snitkovskiy, Dominick Tavitian, Nattanon Bunyatipanon
INFO 370: Data Science
Lavi Aulck & Li Zeng

Works Cited

Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. Journal of Economic Perspective. doi:10.3386/w23089

B.S. Detector. (n.d.). Retrieved October 23, 2017, from http://bsdetector.tech/
Wormald, B. (2014, October 20). Trust Levels of News Sources by Ideological Group. Retrieved October 23, 2017, from http://www.journalism.org/2014/10/21/political-polarization-media-habits/pj_14-10-21_mediapolarization-01/

Iozzio, Corinne. "Reuters Built a bot that can identify real news on Twitter". *Popular Science.* 2016. https://www.popsci.com/artificial-intelligence-identify-real-news-on-twitter-facebook/. Accessed 21 Oct. 2017.

Risdal, M. (2016, November 25). Retrieved October 23, 2017, from https://www.kaggle.com/mrisdal/fake-news/data

Singh, S., Kumar, A., Darbari, H., Singh, L., Rastogi, A., & Jain, S. (2017). Machine translation using deep learning: An overview. *2017 International Conference On Computer, Communications And Electronics (Comptelix)*. http://dx.doi.org/10.1109/comptelix.2017.8003957

Stocking, G. (2017, August 07). Digital News Fact Sheet. Retrieved October 23, 2017, from http://www.journalism.org/fact-sheet/digital-news/