A Machine Learning Analysis of the Features in Deceptive and Credible News

Qi Jia (Jerry) Sun

Abstract

Fake news is a type of pervasive propaganda that spreads misinformation online, taking advantage of social media's extensive reach to manipulate public perception. Over the past two years, fake news has acquired renewed attention due to its perceived impact in the 2016 U.S. Presidential election. Fake news can have severe real-world implications: for instance, in 2016, a man walked into a pizzeria carrying a rifle because he read online that "in this pizzeria, Hillary Clinton was harboring children as sex slaves". This project presents a high accuracy (87%) machine learning classifier that determines the validity of news based off of the word distributions and the linguistic and stylistic differences of the first few sentences of an article. This can help readers identify the validity of an article by looking for specific features in the opening lines and differentiate fake from real news. Using a dataset of 2,107 articles from 30 different websites, this project establishes an understanding of the variations between fake and credible news by examining the model and dataset. This classifier appears to use the differences in word distribution, levels of tone authenticity, and the frequency of adverbs, adjectives, and nouns. The differentiation in the features of these articles can be used to improve future classifiers. This classifier can also be further applied directly to browsers as a Google Chrome extension or as a filter for social media outlets or news websites to reduce the spread of misinformation.

1 Introduction

1.1 Fake News

Fake news is a type of pervasive propaganda that spreads misinformation online, taking advantage of social media's extensive reach to manipulate public perception. Over the past two years, fake news has acquired renewed attention due to its perceived impact in the 2016 U.S. Presidential election. Fake news can have severe real-world implications: for instance, in 2016, a man walked into a pizzeria carrying a rifle because he read online that "in this pizzeria, Hillary Clinton was harboring children as sex slaves". [10]

1.2 Machine Learning

Machine learning is a branch of artificial intelligence that uses statistical models and algorithms to build a mathematical model that can make predictions on unseen data. Machine learning algorithms have been widely adopted for spam email filtering. The machine learning method for classifying spam emails has proven to be successful. Parallels can be drawn between spam emails and fake news such as grammatical errors, misinformation, limited vocabulary, and their purpose of manipulating the reader's opinion for political and financial success. Thus, applying a similar machine learning method from spam emails to fake news could yield promising results.

1.3 Purpose

This project seeks to find a machine learning classifier that determines the validity of news based off of the word distributions and the specific linguistic and stylistic differences of the first few sentences of an article. This can help readers identify the validity of an article by looking for specific features in the opening lines and differentiate fake from real news.

Using a dataset of 2,107 articles from 30 different websites, this project seeks to establish an understanding of the variations between fake and credible news by examining the model and the dataset. The differentiation in the features of these articles can be used to improve future classifiers. A deeper understanding of the differences between deceptive and credible media will further the collective progress in the battle against fake news.

2 Related Work

In December 2016, the fake news challenge was launched (www.fakenewschallenge.org). Its purpose was to explore how machine learning could combat fake news. The challenge was to classify how much a statement agreed with its headline and the winners of this challenge achieved high accuracy rates. This challenge increased the awareness of machine learning in fake news, and within the past few years, many datasets were prepared for fake news classification such as the L.I.A.R. dataset [22], which categorizes news into 5 categories by level of credibility, Rashkin et al., 2017 dataset [17], a dataset differentiating between satire, propaganda, hoax, or trusted news and the BuzzFeedUSE dataset [9], a collection of veracity labels for Facebook links. Another dataset was from Snopes and Politifact which was manually annotated by Asr et al., 2018 contains a large comprehensive set of fact-checked news articles. [4] These datasets have been the foundation for many machine learning models to detect fake news. Many different machine learning models have been created such as the deep diffusive network model (Zhang et al., 2018) [23] or the Naïve Bayes model (Granik et al., 2017) [7]. However, this existing research used unbalanced data sets (Granik et al., 2017), small data sets (Perez-Rosas et al., 2017) [14], or data sets from a single domain (Zhang et al., 2017). Unbalanced datasets between fake and real news may contain significant differences in subject matter, the number of articles, or the date of publishing, contributing to skewed results. Despite the many machine learning models available, no algorithm fully controlled their lurking variables nor determined the features of fake news. This paper introduces a robust classification model which also aids in distinguishing the characteristics that are essential in the composition of fake news.

3 Methodology

3.1 Data Collection

Due to the data-dependent nature of machine learning classifiers, collecting unbiased data is essential to its success. To eliminate all lurking variables, the datasets of fake and credible articles need to be similar in subject matter, writing style, number of domains and perspectives, and publishing date (2 weeks). This allows the classifier to classify based on word distribution and other linguistic features of the articles. Data used in preparation was obtained from 15 credible [11] and 15 fake news sources [2]. Each credible article was meticulously hand labelled by fact-checking and cross-referencing other verifiable sources. These

sources were labelled as credible by established fact-checkers such as PolitiFact or Snopes. Fake news sources were taken from The Fake News Codex [3], a collection of websites known to publish fake news. Furthermore, to eliminate bias, domains were balanced between right-leaning sites and left-leaning sites. Credible news articles were meticulously hand labelled by fact-checking and cross-referencing other articles labelled real at that time.

On this dataset, credible news was on average 205 words longer than fake news. To counteract this problem, the classification included only the first X words of each article. X began at 30 and was found by incrementing in steps of 30 until the classification accuracy leveled off.

3.2 Text Scraping and Cleaning

Text scraping was done through a Java library called JSoup which scrapes text based on its HTML tag. The contents of the articles were scraped from the HTML paragraph element $< p >$. However, many extraneous words in the paragraph element were also scraped such as the report's location, or the journalist's name. These statements were removed from classification. For the classifier to function at its optimal level, each scraped word from the websites needed to be in their inflectional forms because words with the same root (decided, decide, deciding) should be treated as the same word (decide). [19] To achieve this, the text was stripped of HTTPS, removed of non-alphabetical characters, converted into lowercase, and its words stemmed to their root word. [15]

3.3 Data Preprocessing

The first X words in each of the 2,107 articles was compiled (X $\times$ 2,107 words) into a list (List A). X begins at 30 words and increments by 30 words until the accuracy levels off. This classifier is to be trained based on the word distributions of each article. However, it is only important that high-frequency words appear in the article. For example, if a word appeared once or twice, it should not hold any weight in classification. The minimum number of occurrences of a word required in order to be included in the classification is noted in Table 7. Since chi-square tests statistical analyses were conducted on the optimal data set, words that appeared at least 6 times in List A were chosen to be part of the list (List B) of common words. [1]

Words are divided into two categories: function words and content words. Function words are words that signify grammatical relationships (the, to, a ...) while content words have innate meaning (science, green...). Since function words do not contribute meaning they are not included in the classification. Each word on a list of function words found on Semantic Similarity [18] was removed from List B. After removal, List B contained 1000 common words.

To generate article vectors for classification, a matrix M was created with row number (i) equal to the number of articles (2107) and column number (j) equal to the number of words from List B. The order of the articles was randomized to ensure fair representation of all domains. If the i^{th} article contained the j^{th} word, set the value of M_{ij} =1. Otherwise, M_{ij} =0.. A decision column was added to the end to indicate whether the article is credible (1) or fake (0). For example, a fake article vector may look like $< 1, 1, 1, 0, 1 \ldots .0 >$ and a real article vector may look like $< 1, 1, 0, 1, 0 \ldots .1 >$.

3.4 Classification

In order to balance domains, the data were randomly split into training, validation and testing categories. 60% of the data was used for training, 20% for validation, and the remaining 20% for testing. 21 classification models were trained and tested at 95% variance. The top 3 models (based on their classification performance on the validation set) per word count were tested with the testing set to find the overall best model. Furthermore, the errors made by the classifier were analyzed in order to eliminate mistakes from future classifiers and improve upon their accuracy.

3.5 Feature Analysis

Understanding which parts of the article are most important for classification demonstrates the differentiation between fake news and credible news. In essence, the features of an article are dependent on its words, sentences, and paragraphs. A word's features include length, meaning and a part of speech. A sentence's or paragraph's features are its length, its meaning, its tone (sentiment) and its degree of formality and authenticity. However, their meaning is more complex to analyze, often requiring the use of word vectors. Lastly, the overall word distribution in these articles was also very important. The word's part of speech and the overall sentiment (tone) of the text were extracted with the Stanford CoreNLP library [19] [21]. The degrees of formality and authenticity were extracted using the Linguistic Inquiry

and Word Count (LIWC) and Receptivi API [13]. These features were analyzed, and various statistical tests were performed to find their significance and weight (Table 2, 3, 4, 5, 6).

4 Results

Table 1. Classification Model by Number of Words

Number of Words	Model	Area Under ROC Curve	Recall	F1 Score	Testing Accuracy
30 words	Quadratic SVM	58%	52%	57%	54%
60 words	Bagged Trees EC	83%	75%	75%	77%
90 words	Linear SVM	87%	82%	82%	82%
120 words	Linear SVM	87%	83%	82%	83%
90 words after Error Analysis	Linear SVM	87%	82%	82%	87%
120 words after Error Analysis	Linear SVM	87%	83%	82%	87%

Table 2. Part of Speech Distribution in Real and Fake News

	Z-Score	P-Value
Adjective	2.561	0.0104
Adverb	7.225	<0.00001
Noun	-5.428	<0.00001
Pronoun	1.607	0.108
Verb	0.969	0.336

Table 3. Sentiment Score Distribution in Real and Fake News

	Z-Score	P-Value
Very Positive	0.886	0.375
Positive	0.0307	0.976
Neutral	1.154	0.248
Negative	-1.086	0.278
Very Negative	-0.992	0.321

Table 4. Convention Lengths Distributions in Real and Fake News

	Z-score	P-value
Word Lengths	-0.250	0.802
Sentence Lengths	0.651	0.515.

Table 5. LIWC Summary Variable Distribution in Real and Fake News

	Z-Score	P-Value	Chi-Square
Authenticity	1.863	0.0625	$\chi 2$=630.5, df=576, p-value = 0.0574
Emotional Tone	-1.122	0.262	$\chi 2$=398.3, df=380, p-value = 0.249

Table 6. Word Distributions in Real and Fake News

Chi-Squared Test	p-value < 0.00001
Two Sample Z-Test	p-value = 0.0437
Top 10 Most influential words	Clinton, gun, Obama, city, man, link, Facebook, world, school, America, Trump

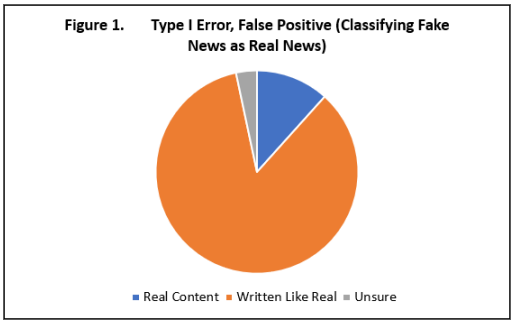
Table 7. Minimum Frequency of Words in List A per Word Count

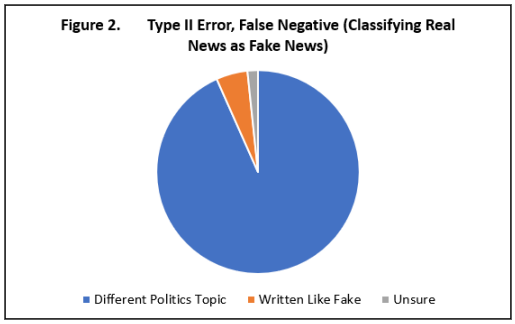
Number of Words	Minimum Frequency of Words
30	4
60	5
90	6
120	6

5 Discussion

5.1 Model Performances

The Linear Support Vector Machine model initially reached a testing accuracy of 83%. Furthermore, the categories of type I and type II errors made by the classifiers were analyzed and split into 3 groups (Figure 1, Figure 2). In the testing set, 200 credible articles were checked to ensure that each article was relevant to the series of topics while simultaneously replacing irrelevant ones with a different article from the same domain and time period. 200 fake articles were also re-fact checked to ensure that the contents were false. In total, 13 fake articles and 21 credible articles were replaced. After this error analysis, both classifying accuracies on the testing set for 90 and 120 words converged at 87%. (Table 1.) This classifier was also used on a Kaggle Fake News Data-set, a comprehensive fact-checked dataset containing over 7500 articles from over 240 domains. The classifier performed at 82.4% on this dataset. This slight decrease in accuracy may be because of the increase in the number of domains.

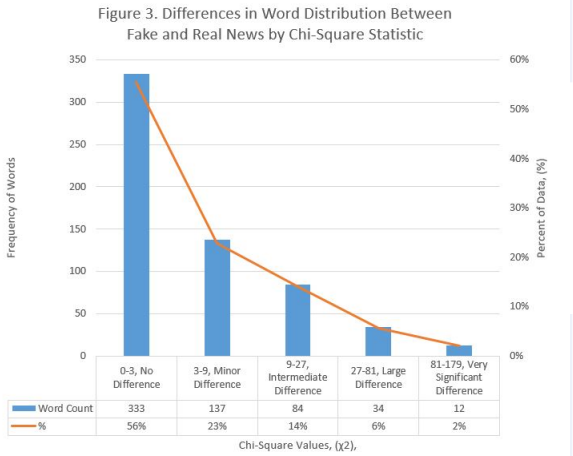


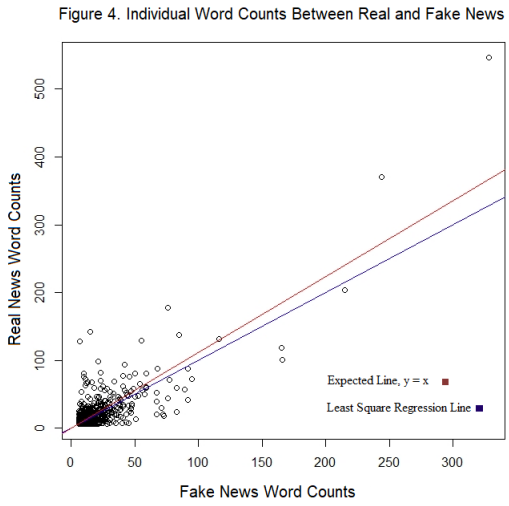


This model (90 words) displayed greater accuracy than both Granik et al. and Zhang et al. even though both utilized skewed data sets. The model classified fake news with an accuracy of 88% and credible news at 85%. Overall the classifiers obtained a recall rate of 82%. This is important as mislabeling credible data could cause the user to trust a deceptive website, increasing their intake of fake news.

5.2 Feature Analysis

Based on this dataset, adverbs were used 40% more in fake news articles ($p < 0.00001$) probably to give emphasis to deceptive information. The higher usage of nouns in credible news could be attributed to credible news presenting more objective information since nouns tend to not hold any emotion. (Table 2.) Furthermore, the p-values of sentiment analysis were not statistically significant. This data was supported by the LIWC summary variable of emotional tone displaying no significant statistical difference between the real and fake news datasets. This contradicts the notion that fake news is emotionally colored. (Table 3.) The differences in word length and sentence length were also statistically insignificant disproving the notion that credible news tends to use more complex language than fake news. (Table 4.) However, LIWC data showed that the authentic tone of fake news was significantly higher than that of credible news. (Table 5.) This deviates from the stereotypical notions of fake and credible news and can be attributed to the fact that fake news puts in more effort to make their content seem authentic. The statistical differences in word distribution were also very significant (Table 6, Figure 3, Figure 4).





The model appears to use all of these distinctions in features to classify fake and real news.

6 Conclusion

This paper presents a high accuracy machine learning classifier to classify the validity of online news articles to be Credible or Fake based on word distributions and other linguistic and stylistic features. The best performing model by overall classification accuracy on the testing set was Linear SVM reaching 87.0%. This model performed better than existing classifiers even with unbalanced datasets such as Granik et al. [7] and Zhang et al. [22]. The optimal number of words (X) in an article to test was found to be 90 words since the accuracy converged after that point. Additionally, by analyzing the model and its features, this project provides insight into the specific features of fake news. This classifier appears to use the differences in word distribution, levels of tone authenticity and the frequency of different parts of speech. In particular, the frequency of adverbs, adjectives, and nouns showed very significant statistical differences between real and fake news.

7 Future Work

This model presented also provides a few limitations which reduce the inferences that can be drawn. This classifier requires large amounts of data to stay updated to each news cycle and produce optimal results. Therefore, finding the minimum amount of data required to still attain high accuracy rates could drastically reduce the time to prepare data for the

classifier. An automated data collection system can be implemented to quickly and cost-effectively collect data. Moreover, this classifier is binary, while news can be partially credible or fake. Thus, adding extra dimensions may present further useful information. This model may be coded into a chrome extension and aid the general public in making better-informed decisions. Further applying a successful classifying filter to news websites or social media outlets can help reduce the large amounts of misinformation circulating the Internet.

8 Acknowledgements

Dr. Maite Tobaoda, a linguistics professor at Simon Fraser University. She has given me the opportunity to take part in a discourse processing lab.
Dr. Fatemeh Torabi Asr, a post-doctorate at Simon Fraser University who has kindly mentored me on data collection and project development. She has given me invaluable advice on extensions to my project.
Dr. Joan Hu, a statistics professor at Simon Fraser University who has given me advice on statistical procedures.
My family, especially my mom and dad for the influence, encouragement, and love that has made this project possible.

References

[1] The Chi-square test of independence. (n.d.). Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900058/

[2] The Fake News Codex. (n.d.). Retrieved from http://www.fakenewscodex.com/

[3] Fake news. (2018, November). Retrieved November 18, 2018, from http://en.wikipedia.org/wiki/Fake_news

[4] Fatemeh Torabi Asr and Maite Taboada (2019) MisInfoText. A collection of news articles, with false and true labels. Dataset.

[5] Function word lists. (2013, November 17). Retrieved from http://semanticsimilarity.wordpress.com/function-word-lists/

[6] Gilda, S. (2017). Evaluating machine learning algorithms for fake news detection. 2017 IEEE 15th Student Conference on Research and Development (SCOReD). doi:10.1109/scored.2017.8305411

[7] Granik, M., Mesyura, V. (2017). Fake news detection using naive Bayes classifier. 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON). doi:10.1109/ukrcon.2017.8100379

[8] How Fake News Goes Viral: A Case Study. (2017, December 22). Retrieved from https://www.nytimes.com/2016/11/20/business/media/how-fake-news-spreads.html

[9] Hyperpartisan Facebook Pages Are Publishing False and Misleading Information at An Alarming Rate. (2016, October 20). Retrieved from https://www.buzzfeednews.com/article/craigsilverman/partisan-fb-pages-analysis

[10] In Washington Pizzeria Attack, Fake News Brought Real Guns. (2018, January 20). Retrieved from https://www.nytimes.com/2016/12/05/business/media/comet-ping-pong-pizza-shooting-fake-news-consequences.html

[11] List of Websites Used for Real News. (2019, March 1). Retrieved from https://drive.google.com/file/d/1bbi4t0MX31TTzVNWibsP0qtsqPd1qHyC/view?usp=sharing

[12] Machine learning. (n.d.). Retrieved March 5, 2019, from https://en.wikipedia.org/wiki/Machine_learning

[13] Pennbaker, J. W., Boyd, R. L., Jordan, K., Blackburn, K. (2015). The Development and Psychometric Properties of LIWC2015. Retrieved from https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf

[14] Perez-Rosas, V., Kleinberg, B., Lefevre, A., Mihalcea, R. (n.d.). Automatic Detection of Fake News. Retrieved from https://arxiv.org/abs/1708.07104

[15] Porter Stemmer Online. (2012, September 7). Retrieved from http://9ol.es/porter_js_demo.html

[16] Precision and recall. (2018, November 3). Retrieved June 3, 2018, from http://en.wikipedia.org/wiki/Precision_and_recall

[17] Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., Choi, Y. (2017). Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. doi:10.18653/v1/d17-1317

[18] Semantic Similarity Function word lists. (2013, November 17). Retrieved from https://semanticsimilarity.wordpress.com/function-word-lists/

[19] Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. Retrieved from https://nlp.stanford.edu/~socherr/EMNLP2013_RNTN.pdf

[20] Stemming and lemmatization. (2009, April 7). Retrieved from http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

[21] Toutanova, K., Klein, D., Manning, C. D., Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03. doi:10.3115/1073445.1073478

[22] Wang, W. Y. (2017). "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). doi:10.18653/v1/p17-2067

[23] Zhang, J., Cui, L., Fu, Y., Gouza, F. B. (n.d.). Fake News Detection with Deep Diffusive Network Model. arXiv.org. Retrieved from https://arxiv.org/abs/1805.08751