

**BT4222**

Mining Web Data for Business Insight

**Group Project: Flagging Fake News**

**Prepared for:**

DR Jovian Lin

**Prepared by:**

Deivapriya RAJENDIRAN A0160989L

KOH Jia Xiang A0161060X

LIAW Sze Wai Sylvia A0161065M

NG Hui Yi Nicole A0116105X

XU Yuwen A0161003A

YAU Chung Yin Jack A0160936B

**Date:**

14 April 2017

# **FLAGGING FAKE NEWS**

Abstract

Fake news refers to deliberately published false stories, misinformation, and propaganda purported by websites to be real news. The consequences of fake news are severe and they include destabilise society, damages to business, and wasting of public’s time and monetary resources. Several countries have called for a tough stance against fake news and failure to remove fake news could render companies to be under investigation and fined. This study attempted to use machine learning techniques to predict the authenticity of news articles and to find out the common characteristics of fake news. There were four major types of predictors, each having potential strategies to tackle the spread of fake news were discussed.

# Introduction

Fake news refers to deliberately published false stories, misinformation, and propaganda purported by websites to be real news. These websites take advantage of social media to drive traffic and, with it, advertising revenue. Though fake news existed due to unethical journalistic practices even before the Internet era, the arrival of social media has led to the proliferation of falsehood by people motivated by profit or sometimes by politics.

Unlike satire news, fake news seeks to mislead rather than to entertain the readers. It takes many forms, while some may be completely fabricated, others are just poorly reported news based on speculations instead of facts. There are still others which take on a misleading narrative to fit an agenda (Schow, 2017). Identifying fake news involves differentiating between its different types, and understanding the underlying structure of fake news articles to separate them from the truth.

The ripple effect of fake news real is detrimental. Mainstream media might pick up the fake news and re-write about it (BBC Trending, 2016). This might lead to other journalists following suit when it reaches a certain amount of traction. Hence, further adding to the authenticity of the fake news.

Social media and search engines have faced heavy criticism for the spread of fake news. Fake news website links are given the same weighting regardless of source, particularly on Facebook, where articles have a potential 1.8 billion audience (Hunt, 2016). Facebook, in its combat against the rise of fake news, plans to flag such stories with the help of users and third-party fact checkers. Facebook also attempts to make website spoofing difficult for these fraudsters thereby reducing the financial incentives to create fake news websites (“Facebook fake news”, 2016). Google had also announced plans to go after the revenue of fake news sites by preventing them from using their advertising networks (“Facebook fake news”, 2016).

# Known Characteristics of Fake News

One feature of fake news was that the headlines of news and domain name must sound as authentic as possible (BBC Trending, 2016). Some fake news websites use website spoofing to imitate online news media websites like ABC news. They adopt the same design, or a similar URL as the target website. URL verification, date of the article, evaluation for reader or writer bias and verification of sources could thus be used to detect fake news (Willingham, 2016).

The first few paragraphs tend to sound legitimate as people stop reading beyond them and the author can spin the rest of the story however possible. Content analysis of the text might also provide signals of the validity of the news since the body of the text might not be coherent.

Another technique used by fake news writers is exploiting readers’ emotions as it confirms people’s belief or prejudice. Sensational headlines or news, especially released during certain major events are of suspect to be investigated further.

Tractions of the news, such as number of likes, replies and readership are potentially good indicators for fake news. This is especially so when the same group of “readers” always respond to the same groups of fake news websites as they are from the same group of companies.

# Business Propositions

The consequences of fake news articles are severe. The spread of fraudulent articles through social media during the 2016 US Presidential elections are said to have affected its outcome. An analysis by Buzzfeed found that in the three months before the election, top performing fake news stories generated more engagement than top stories from major news outlets like the New York Times and CNN (Silverman, 2016). Fake news has also influenced the political discourse of multiple countries including Germany, Indonesia, Philippines, Italy.

Publishing true and unbiased news is the only way to build and maintain the reputation and credibility of the social network. Failing to do so not only damage the trustworthiness of the social media but also on the companies being targeted and the business relationships between them. For instance, Pepsi saw their sentiment score dip on 13th November 2016, a drop of 35% below the average US sentiment score in Q4/2016, and its share price took a significant hit on 10th November 2016, due to a single piece of fake news being published. This eventually became the most negative impactful event for Pepsi in 2016 (“How does Fake News affect corporate reputation?”, 2017; Gupta, 2016).

Apart from reputation risk, social media sites that failed to identify and remove fake news from being posted may find themselves under investigation and fined. The government of Germany has considered imposing a regulation, which will fine social networks up to 500,000 euros per day if the platform fails to remove fake news (Mikelionis, 2016). Countries that already have strong customers’ protection law could impose similar regulations, resulting in more significant penalties for these companies. Besides a heavy fine, the deteriorated business relationship between the impacted companies and the social networks could result in a loss of revenue, causing a double loss for the business.

From the perspective of writers and journalists, fake news writers threaten their jobs and income as a portion of the readers could turn themselves from following genuine news to fake news. With lower readership, bloggers and journalists will struggle to compete with fake news writers, who are willing to generate more articles at lower compensation. Similarly, from the perspective of corporations, genuine news publishers could suffer a loss of revenue as their fake news competitors may generate substandard articles at lower cost while, drawing the attention from the public, gaining higher click rates on their websites by presenting sensationalized article content.

Lastly, readers of fake news could suffer from time and monetary loss. Not only do readers spend more time in identifying the authenticity of the news, they are also at risk of suffering from financial losses if they reacted accordingly to inauthentic news. For example, Bloomberg published a fake press release claiming that French construction giant, Vinci, had uncovered irregularities about hiding losses and sacked their chief finance officer (“Vinci shares plunge on publication of fake press release”, 2016). The shares fell by more than 18% because of the news release. Making serious decisions based on fake news would potentially result in significant loss of the readers and lower confidence towards the social media.

Objectives of Study

With the aforementioned implications of fake news, many people are wondering what data scientists could do to stymie the viral spread of fake news by detecting them. Thus, the goal of the study was to explore various methods of detecting fake news through data analytics with the aim to uncover insights on specific features that were indicative of fake news. In all, the study focused on three areas:

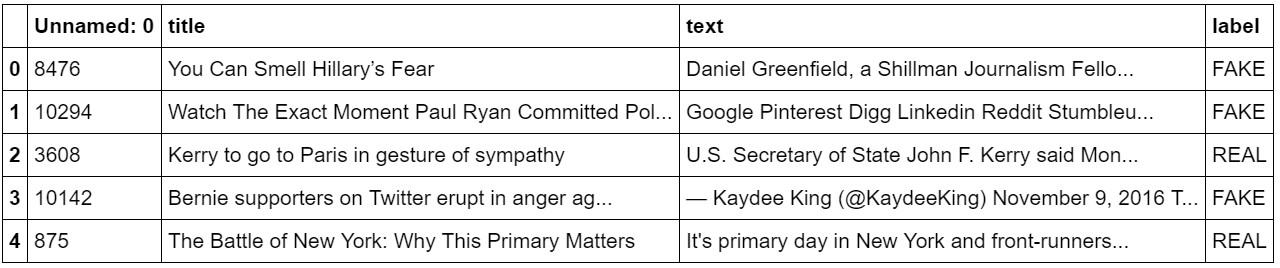
1. Identifying the common characteristics of fake news;
2. Detecting the authenticity of news articles based on the characteristics found;
3. Generating business insights from the common characteristics of fake news.

# Methodology

# Data

The data is made available in an online repository[[1]](#footnote-1) which contains 6,335 articles that were tagged as “REAL” and “FAKE” news. The dataset consisted of US news articles. Table 1 shows a sample of the dataset used in this study.

Table 1: Sample Dataset



As the dataset was textual in nature, data processing[[2]](#footnote-2) was conducted prior to any analyses. The following indicated the procedure of data processing:

1. Removal of stopwords;
2. Deletion of rows with empty title or empty text;
3. Removal of empty lines in the article text;
4. Separation of hyphenated words;
5. Replacement of irregular apostrophes with normal apostrophes;
6. Normalization of unicode data; and
7. Removal of selected punctuations from the dataset, where only exclamation marks and question marks that appeared more than once consecutively were retained.

After the data processing stage, 6,299 records were retained in the dataset. Out of which, 49.7% were labelled as fake and the remaining were labelled as real.

*Data exploration*

An initial data exploration was conducted to investigate the distribution and characteristics of features in the dataset. Such exploration allowed detection of any interesting insights, trends or anomalies in the data. Word Clouds for both real and fake news were created to examine the data quality and occurrences of the words within each dataset as shown in Figure 1. Additional layer of visualization was added to make the word clouds more appealing where the stencil of current American President, Donald Trump was used to represent the fake news and former American President, Barack Obama, to represent the real news. After examining the high frequency words on both classes, the content appeared quite similar, albeit with slightly different frequencies.

Figure 1: Word Clouds of Fake and Real News

# Picture1.png Picture2.png

# Analysis

# The dataset was separated into train and test dataset, where the train data was used to train the various models before using the test data for prediction and comparison of misclassification rate, prediction accuracy and Area Under Curve (AUC) values. Several rounds of modelling were conducted as an iterative process to improve the model results. Naive Bayes (NB), Logistic Regression (LogReg), Decision Tree (Tree), Random Forest (RF), Gradient Boosting Machine (GBM), Support Vector Machine (SVM) and Neural Network (NN) were used for the modelling.

Validation of external news

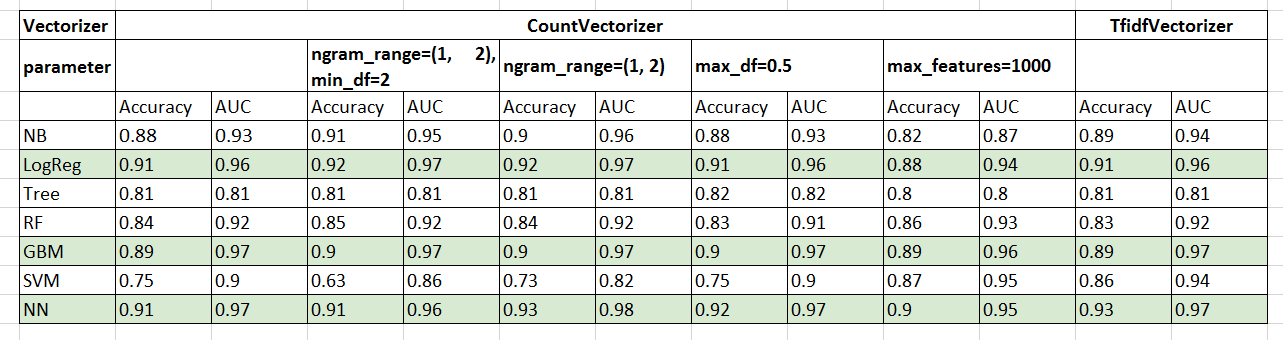
A further validation of a small set of external news[[3]](#footnote-3) (out of sample set) was conducted. The final model would be tested on this small set of external news to lend confidence to the model. One fake news on religion, one fake and one real news on politics from the United States (U.S.) were included. One fake and one real news from Singapore were included.

# Results

First round of modelling

During the first round of modelling, only text vectorization was used to create the features. Both CountVectorizer and TfidfVectorizer with different combinations of parameters were attempted. No parameter tuning was done at this stage to improve the prediction accuracy or AUC values. The Null model with an accuracy of 50.7% was used as a benchmark. Naive Bayes (NB), Logistic Regression (LogReg), Decision Tree (Tree), Random Forest (RF), Gradient Boosting Machine (GBM), Support Vector Machine (SVM) and Neural Network (NN) were used for the statistical analysis. LogReg, GBM and NN were found to outperform other algorithms with an estimated 90% in accuracy and 97% in AUC, as highlighted in green in Table 2. These algorithms were considered as candidate models in the next round of modelling where other engineered features were included.

Table 2: Results of First Round of Modelling using CountVectorizer



Feature engineering

As each record of the raw text had been labelled, the extraction of features within the raw text might identify more insights or characteristics of a fake news. In order to further improve the performance of the model in determining the authenticity of a news piece, new features were created and added to the vectorizer features. With the knowledge of the known characteristics of fake news, four categories (i.e. general features, sentiment analysis, Latent Dirichlet Allocation (LDA) and stance detection) of new features were introduced, which resulted in an addition of 64 features to the dataset. Table 3 shows sample of the 64 features that were added to the dataset.

Considering that fake news articles were more inclined to be written by non-professionals, readability index (i.e. Gunning Fog Index) and the count of punctuations usage such as exclamation mark and consecutive question marks for each news article were tabulated. In addition, fake news might be more intended to trigger emotional responses while real news were more inclined to hold neutral and factual content, hence the count of capital words and the length of the article were added.

Since the label of each news article was known, the use of LDA (Blei, Ng & Jordan, 2003) could be appropriate in determining the features of the topics. To achieve that, further data processing was done to generate the optimal number of topics so as to obtain the probabilities of the news articles for each topic.

Sentiment analysis was done to measure the correlation between headline sentiment and article authenticity. One hypothesis that surfaced was that clickbait articles (containing fake content) rely on more emotive headlines than mainstream news to garner clicks, hence, extreme headline sentiment might be indicative of unreliable article content.

Furthermore, stance detection between the news headline and the article content could be important features. In general, news headline and article content could either agree or disagree with each other on a topic, or discuss on the same topic but not holding specific opinions or unrelated with each other at all (Fake News Challenge). Each scenario may indicate certain probability of a news article being fake or real. Thus, features estimating the relative perspective (or stance) of the headline and content were created.

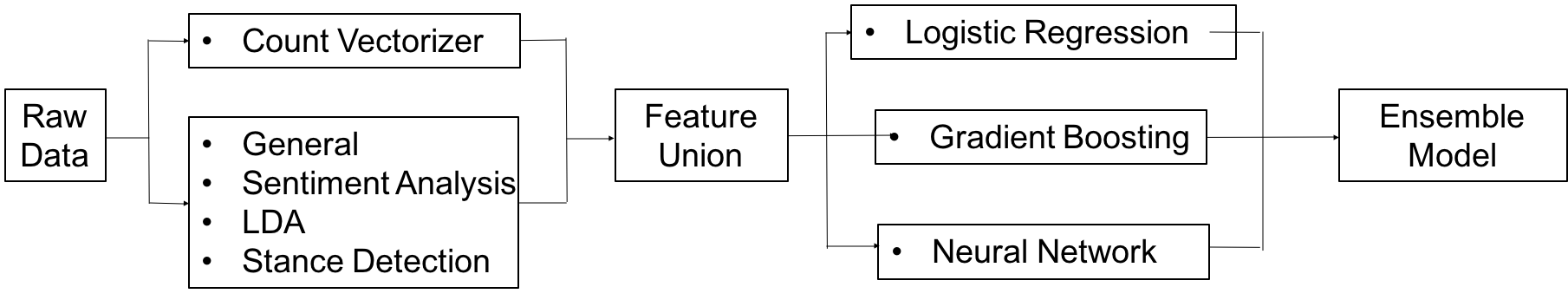
Table 3: Sample of Addition Features

|  |  |  |  |
| --- | --- | --- | --- |
| **General** | **Sentiment Analysis** | **LDA** | **Stance Detection** |
| * Number of words in the news text * Readability index * Number of capital words * Number of exclamation marks * Number of question marks | * Polarity of headlines & text * Subjectivity of headlines & text * AFINN scores of headlines & text | * Topic probabilities, where optimal k (i.e. 4) was determined using the elbow method | * Overlap of tokens between headline and text * Presence of refuting words in headline * Number of refuting words in headline and text * Count of headline tokens’ appearance in the text, n\_grams and char\_grams |

Second round of modelling

In this stage, the aim was to develop a complete pipeline process from feature engineering to prediction. The essential components towards an optimized solution were feature union, pipeline, parameter tuning and model ensembling, as shown in Figure 2. The use of CountVectorizer to generate the document-term matrix and the new curated features were combined using feature union, leading to 5,787 features in total. Pipeline models for Log Reg, GBM and NN were implemented and optimized by applying grid search cross validation to tune the respective parameters. An ensembled model of the three algorithms were implemented with an aim to achieve better prediction performance.

Figure 2: Pipeline



The results of the second round of modelling is shown in Table 4. The accuracy of the model was found to be 0.93 and the AUC value was 0.98. The ensemble model was effective in predicting the authenticity of a news article.

Table 4: Results of Second Round of Modelling

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Models** | **Parameters** | **Accuracy** | **AUC** | **Training time** |
| LogReg | C = 0.1 | 0.93 | 0.97 | 2min 8s |
| GBM | N\_estimators = 100 | 0.91 | 0.97 | 4h 1min 16s |
| NN | Hidden\_layer\_sizes = 100 | 0.93 | 0.98 | 1h 17min 44s |
| Ensemble | - | 0.93 | 0.98 | - |

Table 5 shows the significant predictors. It was found that the shorter the document, the higher the readability index and more capitalised words and exclamation marks, the more likely it is a piece of fake news. Table 6 shows the mean length of news, mean number of capitalised words and exclamation marks. Figure 3 shows the box plot of the readability index, indicating that the variance of readability for fake news was larger.

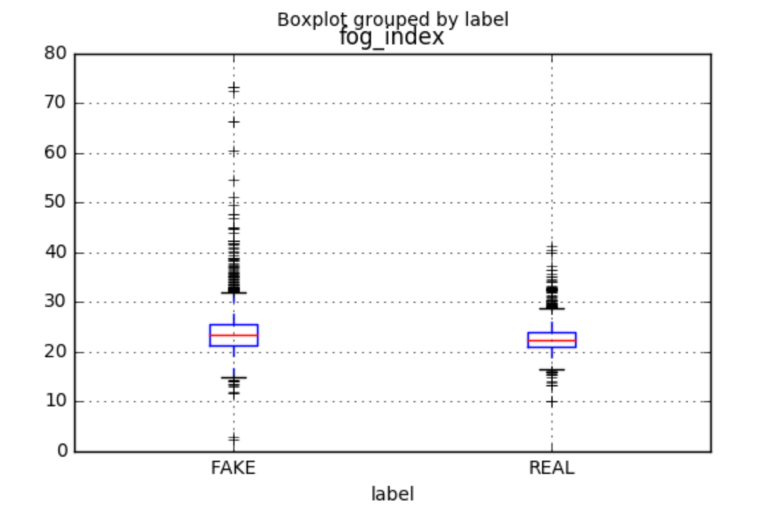
Table 5: Significant Predictors

|  |  |  |  |
| --- | --- | --- | --- |
| **General** | **Sentiment Analysis** | **LDA[[4]](#footnote-4)** | **Stance Detection** |
| * Number of words in the news text * Readability index * Number of capital words * Number of exclamation marks | * Subjectivity of news text * AFINN scores of headlines & text | * Topic probabilities | * Presence of refuting words in headline * Overlap of tokens between headline and text * 6 n\_grams count of tokens’ appearance in news text |

Table 6: Mean Statistics of General Category of Predictors and AFINN Scores of Title and News Text

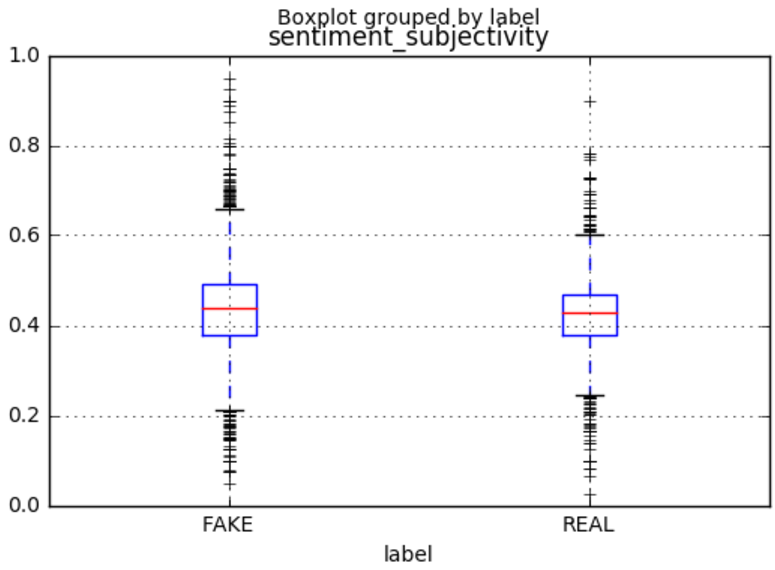
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Mean Length of News Article | Mean Number of Capitalised Words | Mean Number of Exclamation Marks | AFINN Score of Title | AFINN Score of News Text |
| Fake | 4,168 | 6.7 | 0.8 | -0.6 | -8.2 |
| Real | 5,292 | 4.8 | 0.3 | -0.3 | 1.3 |

Figure 3: Boxplot for Readability Index



Fake news had negative sentiment content as compared to real news, which were positive. Fakes news also had a more negative headline than real news and they were more subjective as well. Table 6 shows the mean sentiment scores for title and news text. Figure 4 shows the boxplot of the subjectivity of news text, where fake news had a larger variance and bigger mean subjectivity score.

Figure 4: Boxplot for Subjectivity of News Text



With regard to the features related to stance detection, presence of refuting words such as “deny”, “false” and “despite” in both the headlines and news text, overlap of words between the headline and news text, and six ngrams of word count (i.e. using six words consecutively) in news text were more likely to predict fake news. The results suggested that using refuting words might be attract more readers as people tend to pay more attention to negativity. Also, reading news articles in a more holistic context, that is reading more words, was more important to detect fake news.

Validation of external news

# The final ensemble model correctly classified all the news from U.S. However, the model classified both the Singapore real news as fake.

# Discussion

With the findings generated from the models, it is desired that it can be used in alleviating the issues with fake news by highlighting potential signals on the authenticity of a news piece. This would help readers to mitigate their response if they were had prior knowledge of the characteristics of a fake news. A research study revealed that if readers were pre-warned the genuineness of a piece of news, people could make a better choice in determining the authenticity of a news (Horowitz, 2017).

With the knowledge of the predictors of fake news, educating and raising awareness in the public could be done to help people understand what fake news looked like. For instance, a simple infographic of predictors as a form of communication could be published to educate the public[[5]](#footnote-5). It would help them to spot features on a piece of news by further scrutinizing the article to determine the authenticity of the news.

Schools are already teaching students to be discerning when reading news (Holcombe, 2017). With the knowledge of the significant predictors, schools could refine their approach in teaching students to perceive fake articles from the real ones. The different features could be tiered according to their difficulties in detection. For instance, the general features would be the easiest to detect, while the sentiment and stance detection are more difficult in comparison. Younger children could be taught by focusing on the easy generic features like length of article, number of exclamation marks and capitalised words. If they have trouble understanding an article, it may prompt them to question its credibility as fake articles are generally more difficult to comprehend, according to this study’s findings (i.e. higher readability index). As the children progress, schools could proceed to teach them about assessing the emotions of the articles and comparing the headline and news text if they are coherent.

Another application of the study results could be used by social media sites and genuine news publishers to filter out fake news by doing a direct screening of the predictors. It would be most ideal if they could implement the algorithm and provide an authenticity score to each news article. This score can be further enhanced by providing additional scores (using the same unit of measurement) for the different features. Eventually, readers may subconsciously learn to decipher between real and fake news as they would be able to correlate the score with what they have read. As fake news publishers could also provide fake scores, readers would be able to tell if the scores are consistent with what were published.

Overtime as readers are more au fait with spotting fake news, this behaviour could be further reinforced by encouraging them to report fake news or fake news websites. Small monetary rewards could be given out and they are nothing compared to being fined millions of dollars by the authorities. Hence, it is of the companies’ advantage to encourage such vigilante behaviour. An anti-fake news community could also be set up to gather like-minded individuals who recognize that fake news lead to time and monetary loss, to step up against fake news. Such initiatives would benefit the companies and the society as a whole.

Future works

Due to the large feature set and complexity of the models, especially for GBM and NN algorithms, the modelling process was very time consuming. Table 4 refers. For future improvements, sub-setting the dataset or parallel computing could be employed to reduce the time spent.

While it could predict the additional U.S.’s news (i.e. two politics-related and one religion-related), the prediction for Singapore news were less than ideal. Perhaps the modelling for fake news should be localised in different region of the world. In addition, most of the news in the original dataset were politics in nature, so more testing of other types of news would lend confidence to the model.

Fake news websites tend to have many advertisements as they rely on them for revenue. Hence, considering the number of advertisements that were published together in the news articles might be a good feature in determining the authenticity of the news piece. The types of advertisements such as links to other spam websites could also be studied.

Conclusion

With the knowledge of features that could predict fake news, fake news writers would create new ways to get around being detected. In conclusion, continuous machine learning or artificial intelligence process is needed to keep up with the adaptive behaviour of fake news writers and to combat the new features.

# **REFERENCES**

BBC Trending. (2016, Nov 6). The rise and rise of fake news. BBC News. Retrieved from <http://www.bbc.com/news/blogs-trending-37846860>

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3 (Jan), 993-1022. Retrieved from

<http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf>

Bloomberg, J. (2017, Jan 8). Fake news? Big data and artificial intelligence to the rescue. Forbes. Retrieved from <https://www.forbes.com/sites/jasonbloomberg/2017/01/08/fake-news-big-data-and-artificial-intelligence-to-the-rescue/#506205934a30>

Facebook fake news: Zuckerberg details plans to combat problem. (2016, Nov 19). Retrieved from <http://www.bbc.com/news/world-us-canada-38039506>

Fake News Challenge. Retrieved from <http://www.fakenewschallenge.org/>

Gupta, S. (2016, Dec 6). Trump supporters call to boycott Pepsi over comments the CEO never made. CNN. Retrieved from

<http://money.cnn.com/2016/11/16/news/companies/pepsi-fake-news-boycott-trump/>

Holcombe, M. (2017, Mar 29). Reading, writing, fighting fake news. CNN. Retrieved from <http://edition.cnn.com/2017/03/29/health/school-kids-fight-fake-news-trnd/>

Horowitz, K. (2017, Jan 24). With training, we can learn to spot fake news. Mental Floss. Retrieved from

<http://mentalfloss.com/article/91348/training-we-can-learn-spot-fake-news>

How does Fake News affect corporate reputation? (2017, Jan 09). Retrieved from

<http://www.alva-group.com/en/blog/>

Hunt, E, (2016, Dec 17). What is fake news? How to spot it and what you can do to stop it. The Guardian. Retrieved from

<https://www.theguardian.com/media/2016/dec/18/what-is-fake-news-pizzagate>

Mikelionis, L. (2016, Dec 27). Germany considers fining facebook $522,000 per fake news item. Heatstreet. Retrieved from

<https://heatst.com/tech/germany-considers-fining-facebook-522000-per-fake-news-item/>

Schow, A. (2017, Jan 04). The four types of fake news. The Observer. Retrieved from <http://observer.com/2017/01/fake-news-russia-hacking-clinton-loss/>

Silverman, C. (2016, Nov 17). This analysis shows how viral fake election news stories outperformed real news on facebook. BuzzFeed News. Retrieved from

<https://www.buzzfeed.com/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook?utm_term=.gxOJamP1#.vywwDpQg>

Vinci shares plunge on publication of fake press release. (2016, Nov 22). Retrieved from <http://www.bbc.com/news/business-38071474>

Willingham, E. (2016, Nov 28). A scientific approach to distinguishing real from fake news. Forbes. Retrieved from

<http://www.forbes.com/sites/emilywillingham/2016/11/28/a-scientific-approach-to-distinguishing-real-from-fake-news/#3abb43142692>

**APPENDIX A**

Top 10 Words of the Four Topics Derived from Latent Dirichlet Allocation Analysis

|  |  |  |  |
| --- | --- | --- | --- |
| **Topic 0** | **Topic 1** | **Topic 2** | **Topic 3** |
| Trump | People | State | Police |
| Clinton | World | Clinton | Obama |
| Campaign | United | War | Law |
| Party | Time | Russia | Court |
| Republican | Years | Military | House |
| Hillary | Year | FBI | President |
| President | ! | Obama | People |
| Election | Government | President | Congress |
| Sanders | Percent | People | State |
| Donald | System | Iran | Federal |

**APPENDIX B**

Example of Infographic on How to Spot Fake News[[6]](#footnote-6)



1. Please refer to fake\_or\_real\_news.csv for the dataset. It is also available from <https://github.com/GeorgeMcIntire/fake_real_news_dataset>. [↑](#footnote-ref-1)
2. Please refer to codes06.ipynb in the submission for the complete set of codes for data processing and data analysis. [↑](#footnote-ref-2)
3. Please refer to extra\_news.csv for the external news articles. [↑](#footnote-ref-3)
4. Please refer to Appendix A for the top 10 words in the four topics. [↑](#footnote-ref-4)
5. Please refer to Appendix B for an example of Infographic on how to spot fake news. [↑](#footnote-ref-5)
6. Image retrieved from <https://www.ifla.org/publications/node/11174> [↑](#footnote-ref-6)