A study of fake and real news, including detection via word classes and supervised learning techniques

E. Chew
Department of Computer Science
City, University of London
London, UK
Ethan.Xun@city.ac.uk

***Abstract*—Fake news is a phenomenon that has existed well before the introduction of social media. However, now with social media, news spread across communities and jurisdictions at a rapid rate; including fake news. Not only do fake news spread misinformation, the existence of it gives birth to skepticism and doubt when even reading a piece of real news. Within this paper, we investigate the differences between fake and real news and show that subjectivity, usage of proper nouns and word classes like nouns, pronouns, adverbs and verbs, all serve as an integral component in distinguishing fake news from real ones. We then compare two different supervised learning techniques, for fake news detection, built upon two sets of principal components. Model evaluation suggests the use of a logistic regression model built upon three principal components as a classifier with precision and fall-out of 0.7385 and 0.2671, during testing, respectively.**

Keywords—fake news, real news, classification, natural language processing

I. Introduction

Since the founding of Facebook in 2004 and subsequently YouTube and Twitter in 2005 and 2006 respectively, social media has only continued to grow and saturate at a rapid pace. While each platform has its own appeal and serves its own niche, it is fair to say that all social media platforms have at least one thing in common: open communication with little to no resource. This results in an incomprehensible amount of information that gets circulated daily, including news; both real and fake.

To varying degrees, depending on the type and intention behind it, fake news circulates misinformation with potential repercussions (e.g. serious fabrication). In addition, the complexity of distinguishing fake news from real ones promotes doubt and skepticism when consuming any piece of news or media, let alone fake ones.

Fake news is far from a modern discovery but the combination of open communication channels of social media and the inattention (more than malicious intent) of users [1] leads to challenges in the rate of fake news circulation and traceability. Thus, interventions from social media companies, in fake news identification, may serve as an effective way of combatting the spread of it.

This paper looks to compare the similarities and differences between fake and real news, and present models built upon principal components of the text structure (e.g. number of words of each type) and context (e.g. sentiment), to detect fake news. We will compare the results from two supervised learning techniques, namely: Logistic Regression and Decision Tree.

The remaining sections are structured accordingly. Section 2 highlights the analytical questions and dataset. Section 3 then provides context to the dataset, while Section 4 covers the analysis performed (e.g. data preparation and model development). Section 5 evaluates the findings, makes concluding statements on the study and discusses further work.

II. Analytical Questions and Data

A. Analytical and Research Questions

Overall, this study looks to assess the relationship between text structure and context, and the legitimacy of the news article. From there we will also ask whether fake news detection, from the text structure and context of the text, is plausible using supervised learning techniques. Along the way, we will also be asking intermediate questions such as the following:

- What are the most commonly used words for each news type?
- How do the distributions of the proportions of words in each word class type compare between fake and real news?
- What are the most influential features in detecting fake news?

B. Overview of the Data

The dataset falls between 2016 to 2017, focusing on events surrounding the 2016 US presidential elections and events that occurred during that timeframe. It was first sourced from kaggle.com but was originally created by H. Ahmed et al., as part of their study on using n-gram analysis in detecting fake news [2]. The original dataset gathered real news articles from reuters.com and fake news from unreliable websites posted by Politifact [2]. Both datasets, fake and real news, are separated into its own corresponding datasets (i.e. fake news are collated into one dataset while real news are in the other).

Both datasets are laid out as a corpus and the following information, with the corresponding data type, is available for each instance or piece of article:

TABLE I. Variable Summary

Variable	***Description***	***Data Type***
Title	Title of news article	Text
Text	Body of news article	Text
Subject	Subject matter of news article (e.g. politics, world news, etc.)	Categorical
Date	News posting date	Date
Label	Fake or real news tagging	Categorical

III. DATA (MATERIALS)

Given that the datasets are laid out as a corpus, subsequent tokenisation of the corpus will allow for Wordclouds and the derivation of frequency distributions of each word class type (e.g. adjectives, adverbs, nouns, etc.). In addition, the datasets are labeled allowing for supervised learning techniques to be applied and garner views on feature importance. Furthermore, the distribution of fake and real news is balanced as well with 23,523 and 21,417 instances respectively. With imbalanced datasets, conventional accuracy measures may not be applicable and undersampling and oversampling of the majority and minority class respectively may be beneficial [3].

Rubin et al. suggests three types of fake news, namely: serious fabrications, large-scale hoaxes and humorous fakes (e.g. satirical articles) [4]. All of which benefit from a tailored and custom measures and language processing methods for better classification [4]. However, for this study we will assume that the text structure and context between different types of fake news are similar such that all types of fake news can be viewed holistically rather than granularly.

When comparing the news subjects between real and fake news, the news subjects across both are not necessarily identical. For example, the news subjects for the real news only covers "politicsNews" and "worldNews" while the news subjects for the fake news spans other categories like "Government News", "Middle-east" and "US_News". Thus, following a similar argument, we will assume that the text structure and context between different news subjects for real news are similar such that all types of real news can be viewed holistically rather than granularly. A similar assumption is made for the fake news counterpart as well.

Further assumptions made during data preparation, processing and/or derivation will be highlighted in its corresponding section.

IV. ANALYSIS

Below provides an overview of the workflow:

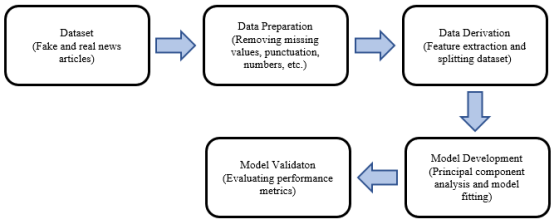

Fig. 1. Workflow Overview

A. Data Preparation

Prior to any tokenisation or feature extraction, the datasets are prepared accordingly:

1. 168 redundant columns are dropped from the fake news dataset.
2. Instances without a title or text are treated as missing values and removed.
3. Fake and real news are then tagged with '0' and '1' respectively and merged into a single dataset.
4. Each article's title and text are merged to a single string to avoid memory issues during tokenisation.
5. Each merged string are parsed through and converted to lowercase and stripped from any punctuation and numbers. This reduces the computational time as irrelevant information is removed.

B. Data Derivation

One challenge of text classification is the large number of dimensions that stem from the number of words used. For this study, we conduct feature extraction by computing the frequency of each word class and sentiment analysis. Feature extraction and processing are performed as follows:

1. Tokenise the merged strings to derive the frequency distribution of word classes for each article. We use the universal part-of-speech tags categories (i.e. verbs, nouns, pronouns, adjectives, adverbs, adpositions, conjunctions, numbers, and particles).
2. Drop any categories that appears less than once per article, on average.
3. Normalise the frequency of each word class by the total frequency of the remaining word classes.
4. Count the number unique words in each article as a proxy for the vocabulary and normalise it similarly. Below provides a statistical overview of the generated features thus far:

TABLE II. STATISTICAL OVERVIEW

	ADJ	ADP	ADV	CONJ	NOUN	DET	VERB	PRON	PRT	vocab
mean	0.107021	0.123994	0.041072	0.025605	0.319893	0.098656	0.198790	0.051716	0.033253	0.572680
std	0.028280	0.023150	0.020461	0.011727	0.052478	0.024454	0.031618	0.025457	0.014338	0.108784
25%	0.089744	0.112026	0.027650	0.018805	0.287129	0.086022	0.181495	0.034211	0.024896	0.503617
75%	0.120787	0.137168	0.053221	0.032573	0.344978	0.113433	0.217391	0.066754	0.041151	0.617470
min	0.000000	0.000000	0.000000	0.000000	0.095238	0.000000	0.000000	0.000000	0.000000	0.131209
max	0.500000	0.315789	0.285714	0.153846	0.857143	0.321429	0.466667	0.285714	0.214286	1.200000

5. Articles with features that sit more than two standard deviations away from the mean are treated as outliers and removed. We assume that sensical articles, on average, consists of a balanced use of words with fairly extensive vocabulary. Articles with high/low proportions of a single word class or highly extensive/limited vocabulary are assumed nonsensical. 11,061 outliers are removed.
6. Perform sentiment analysis using 'TextBlob' to get context of the article in the form of polarity and sentiment scores. Polarity and subjectivity scores represent how positive/negative and subjective/objective an article is respectively.
7. Split the resulting dataset into training and testing using an 80-20 split, amounting to 12 features and 27,063 and 6,766 articles respectively.

The data processing differs from similar work by H. Ahmed et al. [2]. Noticeably, we do not perform stop word removal (i.e. removing common words like 'a' or 'the') or stemming. Stemming and stop word removal may change the distribution of word classes. For example, applying stemming to the noun "runner", will transform it to the verb

"run". Additionally, removal of stop words like "he" or "she" will reduce the frequency of pronouns.

C. Construction of Models

Before fitting a model, we first get more insights on the differences in text structure and context by plotting Wordclouds and compare the means of each feature across fake and real news. The Wordclouds are found below:

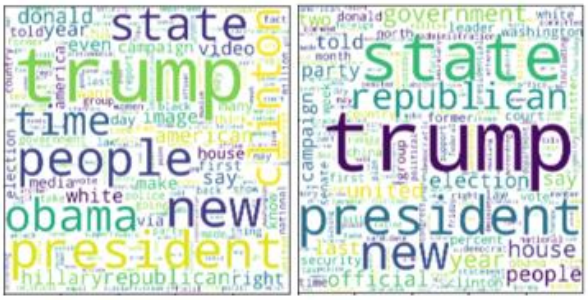

Fig. 2. Wordcloud of fake (left) and real (right) news

Comparing the Wordclouds from Fig.2, both fake and real news share commonly used words like "trump", "president" and "state". This is unsurprising as the dataset is focused around political data but it is noteworthy that both differ in the use of names like "obama", "hillary" and "donald". The plot below provides a comparison of the means across features:

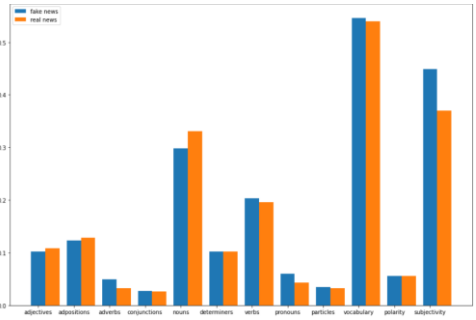

Fig. 3. Comparison of means across features

Fig.3 suggest differing average feature values across fake and real news. Fake news, as compared to real news consists of more adverbs, verbs, and pronouns, and less nouns on average. Additionally, fake news uses more vocabulary and is substantially more subjective than real news on average. The histograms below further reinstate this:

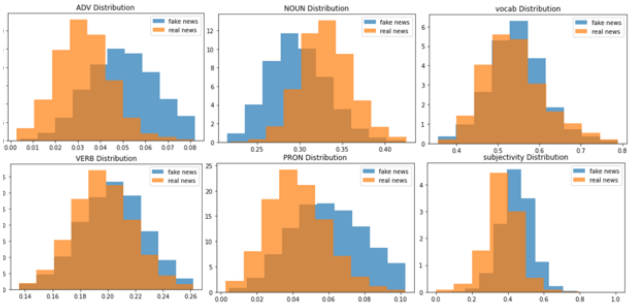

Fig. 4. Distribution of relevant features

The differences observed above increase the plausibility of fake news detection but fitting a model over all 12 features may be challenging. Thus, we employ Principal Component Analysis (PCA) to reduce the dimensionality further to 2/3 principal components. The principal components are a linear combination of our original features and ranked by the variation it explains. Here we choose to compare 2 and 3 principal components for interpretability.

After standardisation, the articles are plotted by the principal components and coloured by their label. Below are the plots for both:

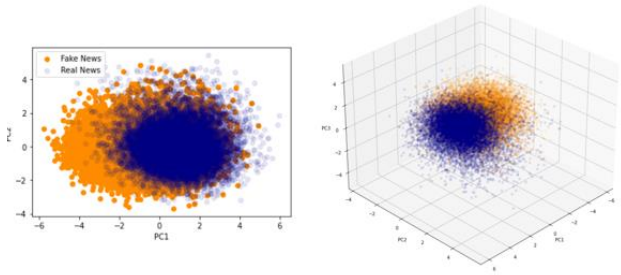

Fig. 5. Scatterplot by 2 (left) and 3 (right) principal components

From the 2-component plot in Fig.5, significant amount of overlap between fake and real news suggest that a third principal component may be beneficial in capturing variation in the z-axis to reduce the number of false positives and false negatives. The 3-component plot then shows marginally more separation when incorporating a third component. Specifically, looking at the explained variance ratio, around 10.82% more variation is explained through the third component.

From our sets of principal components, we look to fit logistic regressions (LR) and decision trees (DT) (4 models in total). Both LR and DT are supervised learning techniques which learns from the labels and features of the dataset. LR models the log-odds as a linear combination of the independent variables while, DT works similarly to a flow-chart with a set of arguments that ultimately lead to leaf nodes. To avoid overfitting, we employ a L2-penalty (i.e. penalty equal to squared magnitude of the coefficients) for LR and depth of DTs are set to 5.

D. Validation of Results

We compare 4 metrics, namely: precision, miss rate, fall-out and the area under the Response Operating Characteristic (ROC) curve (AUC).

We favour models with high precision and AUC with low miss rate and fall-out. Additionally, we are willing to trade a higher miss rate for higher and lower precision and fall-out respectively. These preferences are specific to this study as precision and fall-out captures the false positives which represent the misclassification of fake news.

Below provides a summary of the model performances:

TABLE III. MODEL PERFORMANCE ON TRAINING SET

Model	***Performance Metrics***			
	Precision	***Miss Rate***	***Fall-out***	***AUC***
LR – 2 components	0.7189	0.2555	0.2867	0.7289
LR – 3 components	0.7316	0.2423	0.2738	0.7420
DT – 2 components	0.7176	0.2351	0.2964	0.7342
DT – 3 components	0.7063	0.1641	0.3423	0.7468

Comparing the performance of the 4 models, the 3-component DT prefers positive class predictions most as the (relatively) low miss rate and precision suggest that positive classes are predicted more often. This is further emphasised from the fall-out, as higher fall-out implies higher false positives and/or lower true negatives; both of which stem from misclassification of the negative class.

In addition, the 3-component LR performed the best and falls marginally to the 3-component DT in miss rate and AUC but makes up for it in precision and fall-out. Thus, the 3-component LR is selected as the final model and its performance over the testing set is given below:

TABLE IV. PERFORMANCE METRICS ON TESTING SET

Model	***Performance Metrics***			
	Precision	***Miss Rate***	***Fall-out***	***AUC***
LR – 3 components	0.7385	0.2418	0.2671	0.7455

V. FINDINGS, REFLECTIONS AND FURTHER WORK

A. *Findings*

In this study we've showed the following:

1. Fake news can be differentiated from real news by looking at the types of words used and the sentiment of it. We saw that fake news consist of more adverbs, verbs, pronouns, has more extensive vocabulary and is far more subjective than real news. Additionally, fake news uses fewer nouns but more proper nouns as seen in the Wordclouds (both of which concur with the works done by Horne et al. [5]). Looking at the loadings of the principal components we see a similar argument of feature importance:

TABLE V. ABSOLUTE VALUE OF LOADINGS

Feature	***Components***		
	0	***1***	***2***
Noun	0.4574	0.2450	0.1177
Pronoun	0.4455	0.0382	0.1017
Adverb	0.4121	0.0560	0.2690
Verb	0.4010	0.0866	0.3944
Subjectivity	0.3076	0.0176	0.3072
Adpositions	0.2261	0.4028	0.0923
Adjective	0.2247	0.2911	0.4530
Particles	0.1807	0.1461	0.4478
Determiners	0.1317	0.5877	0.1151
Vocabulary	0.0986	0.5353	0.0661
Polarity	0.0544	0.1275	0.3319
Conjunction	0.0523	0.1053	0.3338

2. Fake news detection by text structure and context is possible using supervised learning techniques and is significantly better than random guessing. Here we fitted a LR model onto the word classes and sentiment, after reducing the dimensionality using PCA, and achieved a testing precision and fall-out of 0.7385 and 0.2671 respectively.

3. For this study, a LR model fitted on 3 components was preferred over 2 components. The overlap between the fake and real news in the 2D space and the marginal separation in the 3D space suggest that a plane in the 3D space may serve as a better decision boundary as compared to a line in the 2D space. Though, a decision boundary in the 2D space is relatively easier to interpret. A visualisation of the decision boundaries are given below:

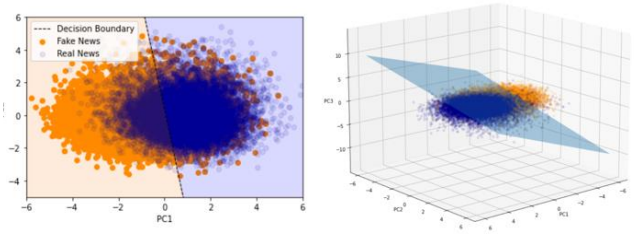

Fig. 6. Decision boundaries in the 2D and 3D space respectively

B. *Reflections and Further Work*

Overall, fairly conclusive results were able to be drawn, with concurrence to similar works done by H. Ahmed et al. [2] and Horne et al. [5]. The dataset was reasonably sized and is practical as it pulled articles from real reliable sources.

However, the dataset scopes in on a specific domain (i.e. news surrounding the 2016 presidential US elections) and may not necessarily be representative of fake and real news from other domains and/or timelines. To test this, it would be interesting to test the performance of this model on other domains that are susceptible to fake news (e.g. COVID-19) and fake news from other years (e.g. 2020 presidential US elections).

Contrastingly, to build an even more focused model, assumptions on the similarity of articles across fake news types and subject matter can be dropped. To do so, further work may include sourcing for fake news of a single type (e.g. satirical news) and strictly focus on a single subject matter (e.g. financial news). Though, sourcing for a strictly focused dataset may prove to be a challenge.

Nonetheless, this study would have benefitted from more granular analysis during feature extraction. Horne et al. suggest that the differences in fake and real news are most distinguishable in the title [5] and so treating the titles and text separately (rather than merging them) may have offered more conclusive results; though with additional computational complexity. Performing sentiment analysis on tokenised sentences of the merged string and averaging it across the sentences (as opposed to performing sentiment analysis on the entire merged string) serves as a potential improvement as well, as doing so may provide a more accurate representation of sentiment of the article.

REFERENCES

[1] G. Pennycook, DG. Rand, "The Psychology of Fake News," *Trends in Cognitive Sciences*, vol. 25, pp. 388-402, May 2021.

[2] H. Ahmed, I. Traore, S. Saad. "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques", 2017, pp.127-138.

[3] N. Chawla, K. Bowyer, L. Hall, W. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," J. Artif. Intell. Res., vol.6, pp.321-357, 2002.

[4] VL. Rubin, Y. Chen, NK. Conroy, "Deception detection for news: Three types of fakes," *Proceedings of the Association of Inforamtion Science and Technology*, vol. 52, pp. 1-4, 2015.

[5] B. Horne, S. Adali. "This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News," in *the 2nd International Workshop on News and Public Opinion at ICWSM*, 2017.

TABLE VI. WORD COUNT

Section	***Word Count***
Abstract	148
Introduction	300
Analytical Questions and Data	271
Data (Materials)	287
Analysis	1000
Findings, Reflections and Further Work	569