A Machine Learning Approach to Identify Fake News

Kamal Shrestha, Jasmin Karki, Divash Ranabhat, Prakash Poudyal

Department of Computer Science and Engineering
Kathmandu University, Nepal
kamalandshrestha@gmail.com,jasminkarki8@gmail.com, rbdiwash@gmail.com,
prakash@ku.edu.np

Abstract. The information age has created several outlets to pave the pathway for public opinion and citizen journalism. With an exponential number of contents being created on a daily basis, a significant part of it includes fake content, or so-called "fake news", usually created with malicious intent. Such an alarming growth of fake news, malicious lies, ineffectiveness of fact-checking and resilience of populist propaganda, demands a system that classifies it to prevent public deceiving and maintain ethical journalism. A promising solution that has come up recently is to use machine learning algorithms to detect patterns in the circulated news that will aid in filtering out the fake content. On this note, we have developed a classification model using the lexical and semantic features extracted from news articles and its sources. Naive Bayes, Support Vector Machine, Logistic Regression, and k-NN models were used and the results were compared to determine the best one among them. Based on precision, recall and f_1 score, k-NN and Logistic Regression gave the most promising results. This is quite inspiring and significant to what was previously developed with similar techniques.

Keywords: Fake News · Machine Learning · NLP · Word Embeddings

1 Introduction

With the development of technology, the number of online news in various social media platforms and news portals has drastically escalated, as it allows an efficient and cheaper way to spread the news in comparison to what was possible with newspapers, TV, and radio. This paradigm of public opinion and citizen journalism not only allows content creators to publish authenticated news but also provides ample opportunity for some of them to deceive the public with "fake news". Any forms of misleading content, false connection, hoaxes, and satire can be classified as fake news. There are several reasons for creating fake news which may be to grab readers' attention, to acquire economic benefits with ad revenues or even to gain political preference. "Anxiousness and curiosity has the potential to spread fake news more quickly than the actual news itself "[22]. Despite the intentions, its alarming growth and spread presents an immediate threat and the need to minimize the negative impacts on the society and people.

Recently, Fake News has been popular since the US Presidential Election Campaign in 2016 for influencing a number of voters suggesting that there might never be a free election in the world again. Fake news has the power of getting people not to believe real things by changing their perspective of interpreting the real news[21]. It has even changed the stock indexes. In 2013, $130 billion in stock value was wiped out in a matter of minutes following an associated press(AP) tweet about an "explosion"that injured Barack Obama[17]. Such unfortunate incidents could have been easily prevented if there were proper filters to recognize what's being circulated to the public, for which many researches have been carried out to find the best possible ways to detect fake news on social media and news portals.

Understanding the need and importance of flushing out fake news from online portals, we propose a classification model to recognize fake news among the many available. Publicly available Kaggle Fake News dataset[18] was used in feature extraction, training and testing the model. Lexical and semantic features of the texts were extracted using POS Tagging, CountVectorizer, term frequency-inverse document frequency (TF-IDF), word2vec and doc2vec from each news which was then used to create the model using Naive Bayes, Linear Support Vector Machine, k- nearest neighbor and Logistics Regression. For the evaluation precision,recall and f-measure were used.

The rest of the paper is organized as follows: Section 2 presents the state of the art on existing techniques applied to classify fake news. Section 3 explains the dataset used for developing our model. Section 4 represents the proposed approach for classification and presents the predictions and evaluations obtained.Section 5 concludes the paper with directions for future works.

2 State of the Art

The evolution of fake news dates back to 13^{th} century BC, when Rameses the Great spread lies portraying the "Battle of Kadesh"as a stunning victory for the Egyptians in which he depicted scenes of himself smiting his foes during the battle on the walls of nearly all temples [23].The spread of fake news started from early sixth century AD by Procopius of Caesarea to discredit the Emperor Justinian in his disclosure called "Secret History"[7]. Some of the approaches that are relevent to this work have been explored.

Khan *et al.*[10] obtained an accuracy of 0.67 on lexical and sentiment features with LSVM whereas Singh *et al.* [19] obtained an accuracy of 0.87 with linguistic features from randomly selected 345 articles of the Kaggle dataset, while Fan [8] obtained the accuracy of 0.896 and Ahmed *et al.* [5] obtained the accuracy of 0.89 on bag of words with LSVM.

Windsor *et al.* [24] compared real news and fake news and proposed the former to have more syntactic complexity, concrete words, deep cohesion and honest language while the latter to have more abstract words, narrativity, referential cohesion and deceptive language. They used Linguistic Inquiry and Word Count (LIWC) features of news headlines and found 68 out of 93 measures, singular

value decomposition (SVD), t-Stochastic Neighbor Embedding (t-SNE) to meet their expectations. Similarly, O'Brien *et al.* [15] were able to identify the language patterns used in fake news articles,responsible for classifying each article using purely text-based approaches.Their testing set of 4,000 articles that was choosen randomly obtained an accuracy of 93.5% with the deflection of 0.2.
Bajaj [6] built a classifier that is able to predict the truthfulness of a piece of news with NLP perspective. The author obtained a precision 0.96, recall 0.49, and f_1 score 0.65 with Logistic Regression and also concluded that RNN with GRUs performed well that one equipped with LSTMs.
Urja [11] used the LIAR dataset and extracted linguistic features like sentiment, subjectivity, number of punctuation marks, quotes, definite articles and soon. The author used algorithms like Baseline, ExtraTrees, Random Forest, Adaboost, Gradient Boosting, SVM and Logistic Regression and obtained the accuracy of 49.03% with GradientBoosting on unigrams, 50.16% with Logistic Regression on POS trigrams and 77.57% with Extra Trees on POS bigrams. She found unigrams, POS tag sequences, punctuation and generality as most determining features to distinguish fake and real news.

3 Data Sets

A collection of fake news was acquired from Kaggle [18] that contained approximately 13000 individual entries flagged bullshit using BS Detector gathered from 244 different websites by Daniel Sieradski[16]. These individual entries were added to a collection of approximately 10000 real news collected from 134 notable sources monitored by News API [1] using application program interface(API) from each of those news sources in a period of 10 days. A combined balanced dataset of 23000 entries was prepared using the collected entries for further data processing (which got reduced to nearly 16261 of balanced data entries after removing entries with missing values). The results of the combined features of lexical and semantic for this dataset are also explained in the section 6.

4 Proposed Approach

The concern for fake and misleading news is global.To address this concern this system is proposed with the objective to develop a classification model that analyses news content, source, title, body, its connectivity, and extracts meaningful information, which might help the model to make predictions. In this approach, we extracted a number of features using news title and news body that might contribute to our prediction. The proposed features were classified into two categories, Lexical and Semantic with an addition of features from word2vec and doc2vec separately. The corresponding features were then trained and tested using the following machine learning classification models and the best was selected to make further predictions: Gaussian Naive Bayes(GNB)[13], Support Vector

Machine (LSVM)[4], Logistic Regression (LR)[25] and k-Nearest Neighbour (k-NN)[12].

Lexical Features

As per the potential of lexical features like noun count, capitalized words, punctuation count suggested by [10] and [11] with an addition of a number of our own potentially contributing features like incorrect use of spaces (after comma, full stop), sentence to word count, incorrect use of punctuation to prepare the combined lexical features.

Table 1 and 2 demonstrates the precision, recall and f_1 score values for four different classification algorithms based on the lexical features. The highest f_1 score for each feature with respective algorithm is marked bold and underlined.

Features	Naive Bayes			SVM			k-NN			LR		
	Pre	Rec	f_1	Pre	Rec	f_1	Pre	Rec	f_1	Pre	Rec	f_1
1	0.671	0.663	**0.661**	0.661	0.660	0.661	0.592	0.601	0.591	0.661	0.662	0.661
2	0.691	0.642	0.623	0.691	0.693	0.692	0.682	0.671	0.672	0.691	0.694	**0.692**
3	0.302	0.554	0.394	0.303	0.549	0.394	0.512	0.473	**0.399**	0.302	0.554	0.390
4	0.462	0.539	0.410	0.302	0.549	0.391	0.542	0.541	**0.542**	0.304	0.549	0.389
5	0.303	0.552	0.392	0.304	0.547	0.389	0.303	0.549	0.392	0.304	0.553	**0.392**
6	0.452	0.553	0.392	0.453	0.554	**0.392**	0.451	0.553	0.392	0.454	0.552	0.391
7	0.304	0.552	0.391	0.304	0.553	0.391	0.302	0.553	**0.392**	0.299	0.548	0.391
8	0.304	0.552	0.392	0.302	0.553	0.383	0.679	0.459	**0.392**	0.299	0.551	0.389
9	0.612	0.557	**0.411**	0.613	0.557	0.411	0.201	0.453	0.283	0.613	0.558	0.410

1)Noun Count 2)Capitalized Words 3) No of Punctuation 4) Sentence to word Ratio 5)Incorrect use of punctuation 6)Incorrect spaces between words 7)Incorrect spaces after comma 8)Incorrect spaces after full stop 9)Incorrect full stop at the end of title
SVM=Support Vector Machine, k-NN=K-Nearest Neighbour, LR=Logistic Regression, Pre=Precision, Rec=Recall

Table 1: Lexical Features Analysis of News Title

The results obtained from lexical analysis of the news title are presented in Table 1. Noun Count and Capitalized Word Count highly influenced in determining the result for News Title. Logistic Regression obtained the highest f_1 score of 0.692 with 'Word Count"while k-NN was consistent in obtaining significant f_1 scores for four out of nine features. Feature 5,6,7 and 8 obtained the f_1 score of 0.392 showing minimum correlation with the class value. This clearly indicates that those features are pretty irrelevant and can be easily ignored when it comes

to News Title.

Features	Naive Bayes			SVM			k-NN			LR		
	Pre	Rec	f_1	Pre	Rec	f_1	Pre	Rec	f_1	Pre	Rec	f_1
1	0.921	0.904	**0.912**	0.912	0.911	0.912	0.909	0.911	0.908	0.909	0.921	0.910
2	0.911	0.910	0.910	0.914	0.923	0.913	0.921	0.919	**0.922**	0.921	0.912	0.913
3	0.932	0.919	0.921	0.934	0.923	0.922	0.928	0.932	**0.929**	0.929	0.921	0.918
4	0.562	0.563	0.462	0.562	0.562	0.462	0.912	0.914	**0.891**	0.552	0.554	0.411
5	0.302	0.553	0.391	0.303	0.549	0.388	0.304	0.553	0.390	0.304	0.547	**0.392**
6	0.721	0.484	0.344	0.302	0.549	0.388	0.304	0.552	**0.393**	0.303	0.459	0.389
7	0.303	0.549	0.391	0.298	0.548	0.389	0.683	0.459	0.298	0.298	0.552	**0.392**
8	0.842	0.753	0.742	0.722	0.719	0.723	0.723	0.721	0.719	0.812	0.771	**0.773**

Indexes are same as used in Table 1

Table 2: Lexical Features Analysis of News Body

Similarly, The results obtained from lexical analysis of the news body is presented in Table 2. k-NN obtained the highest f_1 score of 0.929 with "No of Punctuation Count". In comparison to the results from News Title, News Body seemed to have significantly higher values for each feature, which is mainly because of the increased amount of corpus text corresponding to significant numerical values.
Comparing Table 1 and 2, the low f_1 score of labelled features 5, 6 and 7 as compared to other features doesn't necessarily invalidate its significance but shows less correlation in the corpus. Nevertheless, these features are one of the most prominent grammatical syntax in English language and should always be considered.

Semantic Features:
The features like cosine similarity using count vectorizer & TF-IDF [20], Sentiment Analysis(Polarity and Subjectivity), Source Credibility, Word2vec, and Doc2vec were used. Unlike lexical features that reflect the form, style and correctness of writing, semantic features reflect the correlation, similarity and sentiment between title and body.
Cosine Similarity uses previously obtained vectorical forms of news texts in the form of TF-IDF to calculate text and document similarity. Comparisons were made in following fashion for both news title and news body.

1. Between the News Title and News Body

2. News Title of a given news and News Title of Several pre-defined real news (in our context at minimum of 10 real news were used)
3. Same News Title and News Title of Several pre-defined fake news

Cosine similarity values were expressed into two different forms: Similarity Ratio (Ratio of total similarity of a specific news compared with both fake and real news, with the total availability of news, gives and idea about how similar it is to real or fake news) and Numerical Similarity (Ratio of total number of similar news to the total number of news, gives us an idea about how many news it is similar with).
Source Credibility is vital and crucial to categorize our news sources in terms of their credibility to being an authenticate news source. In lack of any proper rankings, measurement criteria and source authenticity, we were forced to use the list of biggest publishers on Facebook ranking table provided every month by Facebook[2] as our primary selection list of reputed and notable sources, based on the tallied score of total likes, shares, comments, and reactions for that given month.
Sentiment Analysis was carried out using Text Blob library of standard NLP, which measures the sentiment of the given sentence in terms of two defining parameters:- Polarity and Subjectivity.[14]
The numerical values obtained from these semantic features are now modelled in to given algorithms and tabulated in table 3 and 4 for News Title and News Body respectively:

Features			**Naive Bayes**			**SVM**			**k-NN**			**LR**		
			Pre	Rec	f_1	Pre	Rec	f_1	Pre	Rec	f_1	Pre	Rec	f_1
1	NH & NB		0.601	0.385	0.470	0.571	0.511	0.539	0.580	0.539	**0.559**	0.569	0.514	0.540
	RN	SR	0.550	0.409	0.469	0.540	0.488	0.513	0.507	0.669	**0.577**	0.540	0.488	0.513
		NS	0.535	0.617	**0.573**	0.545	0.419	0.474	0.526	0.520	0.523	0.545	0.419	0.474
	FN	SR	0.539	0.757	**0.630**	0.544	0.617	0.578	0.567	0.673	0.616	0.545	0.615	0.578
		NS	0.539	0.663	**0.595**	0.536	0.496	0.516	0.493	0.664	0.565	0.540	0.496	0.516
2	Polarity		0.528	0.199	0.289	0.576	0.754	0.606	0.517	0.721	**0.608**	0.507	0.754	0.606
	Subjectivity		0.509	0.291	0.310	0.494	0.453	**0.472**	0.514	0.307	0.385	0.494	0.453	0.427
3			0.499	1.000	0.666	0.499	1.000	0.666	0.499	1.000	**0.666**	0.499	1.000	0.666

1)Cosine Similarity 2)Sentiment Analysis 3)Source Credibility
NH= News Head, NB= News Body, RN=Real News, FN= Fake News, SR= Similarity Ratio, NS= Numerical Similarity

Table 3: Semantic Features Analysis of News Title

As seen in Table 3, Apart from k-NN which was pretty consistent till now, Naive Bayes algorithm also performed well . The f_1 score of 0.630 attained for-Similarity Ratio of Cosine Similarity"value compares an individual news item to a number of pre-defined fake news items. Source Credibility turned out to be the most determining feature with an f_1 score of 0.666. Although it is unlikely for reliable sources to publish fake content, the score of only 0.666 is as a result of some instances of news even from reliable sources marked as bullshit by BS Detector.

Features			**Naive Bayes**			**SVM**			**k-NN**			**LR**		
			Pre	Rec	f_1	Pre	Rec	f_1	Pre	Rec	f_1	Pre	Rec	f_1
1	RN	SR	0.7679	0.862	0.812	0.776	0.814	0.794	0.790	0.844	**0.816**	0.776	0.816	0.796
		NS	0.508	0.985	**0.670**	0.508	0.985	0.669	0.115	0.001	0.002	0.508	0.977	0.669
	FN	SR	0.801	0.900	0.847	0.805	0.882	0.842	0.837	0.870	**0.853**	0.801	0.874	0.839
		NS	0.501	0.971	0.661	0.501	0.971	0.661	0.625	0.003	0.007	0.501	0.971	0.661
2	Polarity		0.566	0.876	0.688	0.536	0.902	0.672	0.755	0.702	**0.727**	0.535	0.903	0.672
	Subjectivity		0.601	0.894	0.719	0.580	0.689	0.629	0.773	0.721	**0.746**	0.580	0.689	0.630

Indexes are same as in Table 3

Table 4: Semantic Features Analysis of News Body

In Table 4, k-NN performed well for five out of eight features with highest being 0.853 for "Similarity Ratio of cosine similarity value". The highest value of f_1 score for semantic analysis of News Title is 0.666 where as the highest for News Body is 0.853, which is far superior and better result.Similarly, Cosine similarity between news title and news body obtained an f_1 score of 0.559. In comparison to results in table 4,it is significantly low, indicating that the dataset contains instances of news with proper correlation between news title with its body.Comparing the structure and similarity of an individual news with a number of predefined fake and real news, expressed by similarity ratio, obtained inspiring f_1 score of 0.816 and 0.853 respectively. This clearly expresses the idea that a collection of fake news share some common characteristics, be it grammatical error or form of writing. It can also be said the other way around, that real news stories share a common characteristics . Either way, it aids in classification.

*Word2Vec:*To the existing dataset describes in section 3, we appended Google's Standard Data Dumps [3] to develop a word2vec model. *Word MoversDistance(WMD)* [9] was calculated between two sentences (in our context the news title and news body).Calculations were carried in similar fashion to cosine similarity for each news title and news body.

*Doc2Vec:*Similar to word2vec, documents to vector were also modelled out using the available texts from news descriptions and headlines from datasets and iterated to develop a model in 10 epochs. Also, In addition to calculation of cosine similarity between news title and news body using vectorized values from doc2vec, the actual vectorized doc2vec value of 25 dimensions each, of the sentence was also used in prediction.

Features	Naive Bayes			SVM			k-NN			LR		
	Pre	Rec	f_1	Pre	Rec	f_1	Pre	Rec	f_1	Pre	Rec	f_1
Lexical	0.990	0.876	0.930	0.987	0.865	0.922	0.984	0.887	0.933	0.974	0.926	**0.950**
Semantic	0.570	0.980	0.720	0.848	0.853	0.851	0.851	0.853	**0.852**	0.832	0.847	0.839
Doc2vec	0.952	0.890	0.920	0.967	0.884	0.924	0.962	0.890	**0.925**	0.954	0.889	0.920
Word2vec	0.738	0.811	0.773	0.640	0.794	0.709	0.636	0.827	0.719	0.805	0.777	**0.791**
Combined	0.991	0.893	0.947	0.981	0.915	0.947	0.973	0.939	**0.956**	0.979	0.930	0.954

Table 5: Combined Features Analysis

Table 5 demonstrates the combined results for each category of features including word2vec, doc2vec and the overall combined. An overall f_1 score of 0.956 was obtained using k-NN which seemed not so surprising because it was the most consistent one as seen in previous tables. With lexical analysis it was pretty clear that fake news instances in our dataset contained a lot of grammatical errors compared to real news instances which gave our classification model a clear distinction between two news. This was solely the reason to obtain an outstanding f_1 score of 0.950 by lexical analysis, which is not far from the highest attained overall. This might certainly question the efforts and calculations needed for other categories. But in case of fake news being well edited grammatically, the current score of 0.950 from lexical analysis is more likely to drop drastically. Hence, it is vital to carry out all the analysis before making predictions.

5 Conclusion and Future Work

The problem of fake news has provoked negative impacts on the society and people's perception towards the technology. So, In this research, we proposed a system that focused on predicting whether the given news is falsified or not, as improvised extension of previous works with appended features. k-NN performed consistently with most of the features in both categories including doc2vec and can be esily used in making further predictions. The overall results that we obtained are quite inspiring and promising as compared to other researches done

on the same topic.
Furthermore, larger datasets could be introduced in this existing model to narrow down the most significant features. We could also implement the concepts of highly computational neural networks to make more accurate predictions.

References

1. News api, https://newsapi.org/sources, accessed: 2019-06-15
2. List of top facebook publishers. https://tubularinsights.com/top-facebook-publishers-march-2019/ (2019)
3. Rare technologies, word2vec google news data dump. https://github.com/RaRe-Technologies (2019)
4. Abe, S.: Support vector machines for pattern classification, vol. 2. Springer (2005)
5. Ahmed, H., Traore, I., Saad, S.: Detecting opinion spams and fake news using text classification. Security and Privacy **1**(1), e9 (2018)
6. Bajaj, S.: The pope has a new baby! fake news detection using deep learning. Tech. rep., Technical Report, Stanford Univ (2017)
7. Burkhardt, J.M.: Combating fake news in the digital age, vol. 53. American Library Association (2017)
8. Fan, C.: Classifying fake news. conniefan. com (2017)
9. Huang, G., Guo, C., Kusner, M.J., Sun, Y., Sha, F., Weinberger, K.Q.: Supervised word mover's distance. In: Advances in Neural Information Processing Systems. pp. 4862–4870 (2016)
10. Khan, J.Y., Khondaker, M., Islam, T., Iqbal, A., Afroz, S.: A benchmark study on machine learning methods for fake news detection. arXiv preprint arXiv:1905.04749 (2019)
11. Khurana, U., Intelligentie, B.O.K.: The linguistic features of fake news headlines and statements. Ph.D. thesis, Masters thesis, University of Amsterdam (2017)
12. Kim, H.J., Tomppo, E.: Model-based prediction error uncertainty estimation for k-nn method. Remote Sensing of Environment **104**(3), 257–263 (2006)
13. Lowd, D., Domingos, P.: Naive bayes models for probability estimation. In: Proceedings of the 22nd international conference on Machine learning. pp. 529–536. ACM (2005)
14. Mejova, Y.: Sentiment analysis: An overview. University of Iowa, Computer Science Department (2009)
15. O'Brien, N., Latessa, S., Evangelopoulos, G., Boix, X.: The language of fake news: Opening the black-box of deep learning based detectors (2018)
16. Petegem, S.V., Brandstetter, S., Maass, R., Hodge, A.M., El-Dasher, B.S., Biener, J., Schmitt, B., Borca, C., Swygenhoven, H.V.: On the microstructure of nanoporous gold: an x-ray diffraction study. Nano letters **9**(3), 1158–1163 (2009)
17. Rapoza, K.: Can fake news impact the stock market? Pridobljeno iz www. forbes. com/sites/kenrapoza/2017/02/26/can-fake-news-impact-the-stock-market/(9. 7. 2018) (2017)
18. Risdal, M.: Getting real about fake news. https://www.kaggle.com/mrisdal/fake-news (2016), accessed: 2019-06-30
19. Singh, D.V., Dasgupta, R., Ghosh, I.: Automated fake news detection using linguistic analysis and machine learning. In: International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS). pp. 1–3 (2017)

20. Singh, P.: Natural language processing. In: Machine Learning with PySpark, pp. 191–218. Springer (2019)
21. Tavernise, S.: As fake news spreads lies, more readers shrug at the truth. https://www.nytimes.com/2016/12/06/us/fake-news-partisan-republican-democrat.html (2016), accessed: 2019-08-25
22. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**(6380), 1146–1151 (2018)
23. Weir, W.: History's Greatest Lies: The Startling Truths Behind World Events Our History Books Got Wrong. Fair Winds Press (2009)
24. Windsor, L.C., Cupit, J.G.: Syntactic, semantic, and topics: The cognitive framework of fake news
25. Wright, R.E.: Logistic regression. (1995)