Comparative Analysis of Machine Learning Algorithms for Fake News Detection

Nikhil Madaan
Birla Institute of Technology, Pilani
Hyderabad, India
nikhilm29.09@gmail.com

Siddharth Dharm
Birla Institute of Technology, Pilani
Hyderabad, India
siddharthdharm@gmail.com

Aruna Malapati
Birla Institute of Technology, Pilani
Hyderabad, India
arunam@hyderabad.bits-pilani.ac.in

ABSTRACT

With a surge in the number of electronic devices connected to the Internet, the propagation of news (both, legitimate and fake) is inevitable. The dataset comprises of 9,400,000 labeled news articles. In this paper, we propose to use readability and linguistic features extracted using NLP (Natural Language Processing) techniques, which help in substantially improving the performance of the classifiers. Our experimental study consists of comparing the performance of eight classification algorithms, thereby suggesting the best algorithm for detecting fake news.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Information systems** → *World Wide Web*.

KEYWORDS

Natural Language Processing, Computational Linguistics, Artificial Intelligence, Deep Learning, Artificial Neural Networks, Recurrent Neural Networks

ACM Reference Format:

Nikhil Madaan, Siddharth Dharm, and Aruna Malapati. 2018. Comparative Analysis of Machine Learning Algorithms for Fake News Detection. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

In the Internet of Things era, no less than twenty billion devices are connected to the Internet. Statistical reports show that this number will continue to increase. The dissemination of information also has become inexpensive and hassle free. Moreover, the internet has made it easy to post content without any restrictions. On account of these reasons, fake news, along with legitimate news, spreads like wildfire.

Several research groups have delved into this problem, however, very few of them have made use of Linguistic properties of the English (natural) language. In this paper, we have developed a model that incorporates an array of linguistic features. These features have deeply impacted our model, and have enabled us to achieve accuracies over 95 per cent. We compare the performance of eight classification algorithms in this paper to highlight various nuances.

Hyderabad '19, Dec 16–17, 2019, Hyderabad, India

ACM ISBN 978-1-4503-9999-9/18/06…$15.00

2 RELATED WORK

Till date, considerable amount of research has been conducted on applying Machine Learning/Deep Learning algorithms for detecting fake news.

Moreover, [Potthast et al. 2018], recently proposed a writing-style approach to identify of fake news articles. They used the Buzzfeed dataset (manually labelled) but were not able to segregate fake news from the genuine news articles with sufficient accuracy.

[Conroy et al. 2015] provided a survey analysis which indicate that the major categories of fake news detection methods are Linguistic Approaches and Network Approaches. (in which network information like message metadata is used) They have proposed a hybrid solution to detect fake news, in which both, linguistic and network data is harnessed.

There exist certain computational approaches for detecting fake news, which have employed satirical news sources and fact-checking websites. As brought out by [Perez-Rosas et al. 2018], the use of satirical sources poses potential drawbacks. Also, [Perez-Rosas et al. 2018] have pointed out that fact-checking websites are generally constrained to a particular domain of interest which make it quite cumbersome to obtain generalised datasets.

3 DATASET

The paper makes use of the following dataset:-

3.1 FakeNewsCorpus

A dataset obtained from the [Szpakowski 2017] , containing 9,408,908 labelled news articles. These articles have been scraped from a curated list of 1001 domains. This corpus is aimed toward training algorithms for detecting fake news. It includes news articles related to a number of 'tags', like fake, reliable, satire, bias, etc. 60,000 articles were selected randomly from the corpus. These articles had their 'tag' attribute either equal to 'fake' or 'reliable'. During random sampling from the corpus, it was ensured that articles belonging to both the categories, were represented equally.

4 FEATURES

Our paper makes use of different combinations of the following features in different algorithms. The features being used are the following:-

4.1 Word Embeddings

The text is being represented using word embeddings[Mikolov et al. 2013a]. A pre-trained Google word2vec model [Mikolov et al.

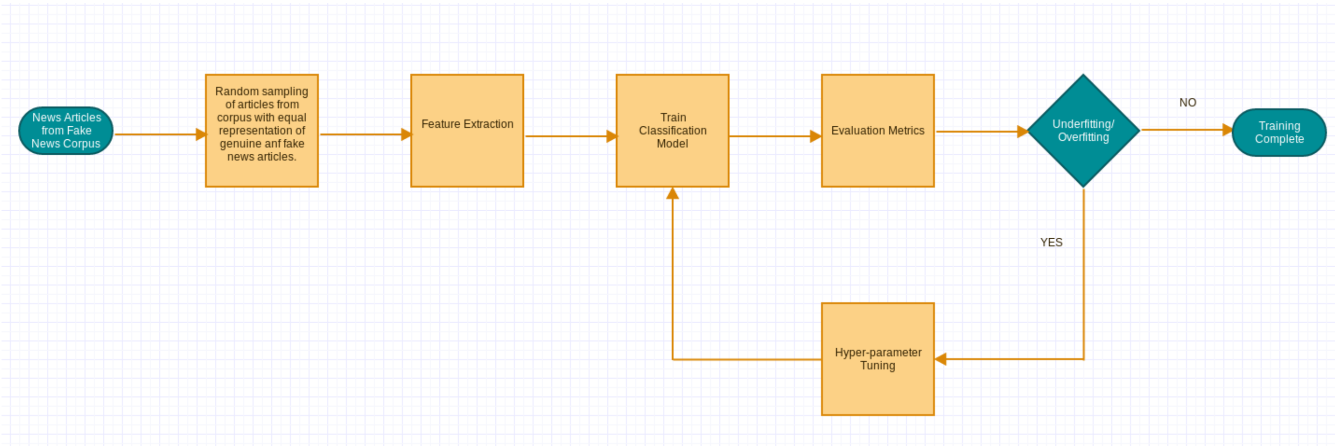

Figure 1: Training Model Architecture

2013b] is being used to get the vectorized representations (of 100 dimensions) for the words. To represent the text, the mean of Word Embeddings (only of those words in the text which are a part of the Word2Vec vocabulary) are taken into account.

4.2 Readability Features

In order to distinguish the fake news from real ones based on the writing style we plan to use several hand-designed text level readability features from a given news text input. We expect that these features will provide the additional discriminative power to the learning algorithm as they take the global features of news into account as compared to simple word based local features generally used. The following readability features have been extracted for classification:-

- Syllable Count:- This returns a count of the number of syllables in a particular text.
- Sentence Count:- This returns a count of the number of sentences in a given text.
- Flesch – Kincaid Grade Level:- This score represents the number of years of education generally required to understand a particular text. The formula from [Eltorai et al. 2015] is as follows:

$$F - KGradeLevel = 0.39\frac{TW}{TST} + 11.8\frac{TS}{TW} - 15.59 \quad (1)$$

  where TW stands for Total words, TST stands for Total sentences and TS stands for Total syllables.
- Flesch Reading Ease:- As the name suggests, this score represents the ease with which a particular text can be read. A higher score indicates that the text is easy to read; a lower score marks passages that are difficult to read. The formula from [Karmakar and Zhu 2010] is as follows:

$$FRE = 206.835 - 1.015\frac{TW}{TST} - 84.6\frac{TS}{TW} \quad (2)$$

  where TW stands for Total words, TST stands for Total sentences and TS stands for Total syllables.
- Automated Readability Index (ARI):- This index is designed to gauge the understandability of a text. The formula from [Eltorai et al. 2015] is as follows:

$$ARI = 4.71\frac{C}{W} + 0.5\frac{W}{S} - 21.43 \quad (3)$$

  where C stands for number of letters and numbers , W stands for number of spaces and S stands for number of sentences.
- Gunning Fog Index:- This index produces an estimate of the number of years of formal education required by a person to understand a given text on the first reading. The formula from [Karmakar and Zhu 2010] is as follows:

$$GFI = 0.4\frac{W}{S} + 40\frac{CW}{W} \quad (4)$$

  where W stands for number of words, S stands for number of sentences and CW stands for number of complex words (words consisting of three or more syllables excluding proper nouns, familiar jargon, or compound words).
- SMOG Grade:- The SMOG (Simple Measure of Gobbledygook) grade is a measure of readability similar to the Gunning Fog Index that estimates the years of education needed to understand a piece of writing. The formula from [Karmakar and Zhu 2010] is as follows:

$$SMOG = 1.043\sqrt{NPS * \frac{30}{NS}} + 3.1291 \quad (5)$$

  where NPS stands for number of polysyllables and NS stands for number of sentences.
- Linsear Write:- This is a readability metric for English text, designed to calculate the readability of technical writing. There is a sequence of steps to be followed to calculate the Linsear Write Score. The steps for computation from [Eltorai et al. 2015] are as follows:

(1) Find a 100-word sample from the given text.
(2) Calculate the number of easy words (defined as two syllables or less) and place a number "1" over each word, even including a, an, the, and other simple words.
(3) Calculate the hard words (defined as three syllables or more) and place a number "3" over each word as pronounced by the dictionary.
(4) Multiply the number of hard words times 3.

(5) Add the two previous numbers together.
(6) Divide that total by the number of sentences.
(7) If the answer is greater than 20, divide by "2" to get the Linsear Write Score.
(8) If the answer is less than or equal to 20, subtract "2" and then divide by "2" to get the score.

- Dale-Chall Readability Score:- This score provides a numeric gauge of the difficulty of comprehension that readers face while reading a given text. The formula from [Dubay 2004] is as follows:

$$DCRS = 15.79\frac{DW}{W} + 0.0496\frac{W}{S} + 3.6365 \tag{6}$$

where DW stands for number of Difficult words, W stands for number of words and S stands for number of sentences.

- The Coleman-Liau Index:- This score is used to gauge the understandability of a text (designed by Meri Coleman and T.L.Liau), this metric can be calculated using the following formula from [Karmakar and Zhu 2010] :

$$CLI = 5.89\,L - 29.5\,S - 15.8 \tag{7}$$

where L stands for average number of letters per hundred words, S stands for average number of sentences per hundred words.

4.3 Parts-of-Speech

The count of following parts of speech have been extracted for classification:-

- Nouns
- Adjectives
- Verbs
- Adverbs

We restrict the Part-Of-Speech(POS) feature to these limited set since they are most informative and generic unlike others POS, like Conjunction and Preposition.

5 METHODOLOGY

For the experimental study, this paper presents a comprehensive analysis of the following algorithms:

- Logistic Regression is a classification algorithm for classifying both discrete and continuous input features into binary class output.
- Random Forest is an ensemble learning method that makes use of a forest of decision trees for building the classification model.
- Support Vector Machines provides us with the ability to obtain a hyperplane in space defined by kernel function, which serves as a decision boundary for classification.
- Artificial Neural Network or simple feed-forward Neural Network, with multiple layers, provide us with the capacity to learn very non-linear decision boundary. But they don't capture the time ordering in data hence less suited for tasks related to Natural Language Processing.

Table 1: Evaluation Metrics of Classification Algorithms

Algorithm	Accuracy	Sensitivity	Specificity	Precision
Logistic Regression	0.9444	0.9397	0.9492	0.9493
Random Forest	0.9481	0.9302	0.9659	0.9645
SVM	0.8947	0.8836	0.9045	0.9026
ANN	0.9574	0.8636	0.9454	0.9416
LSTM(50-D)	0.9721	0.9866	0.9574	0.9587
Bi-LSTM(50-D)	0.9796	0.9895	0.9774	0.9777
LSTM(100-D)	0.9723	0.9682	0.97627	0.97620
Bi-LSTM(100-D)	0.9765	0.9875	0.9654	0.9663
LSTM(200-D)	0.9669	0.9945	0.9388	0.9429
Bi-LSTM(200-D)	0.9828	0.9945	0.9388	0.9429
GRU(50-D)	0.9312	0.9895	0.9249	0.9347
Bi-GRU(50-D)	0.9290	0.9857	0.9380	0.9355
GRU(100-D)	0.9225	0.9790	0.9372	0.9343
Bi-GRU(100-D)	0.9232	0.9744	0.9315	0.9389
GRU(200-D)	0.9222	0.9667	0.9318	0.9451
Bi-GRU(200-D)	0.9258	0.9704	0.9446	0.9457

- LSTM (Long Short-Term Memory) from a family of more general Recurrent Neural Network proposed by [Hochreiter and Schmidhuber 1997], allows us to capture temporal structure in the sentence implicitly.
- Bi-directional LSTM (Long Short-Term Memory) a variant of original LSTM, provides capacity to process of sentence in both forward and backward direction.
- GRU (Gated Recurrent Unit) another instance from family of Recurrent Neural Network proposed by [Chung et al. 2014], as a modification to original LSTM architecture with comparable performance but with lesser parameter.
- Bi-directional GRU (Gated Recurrent Unit) is similar extension of GRU, to provide bi-directional information processing capacity.

For first four algorithms, the input to our model is the feature set described in the earlier sections. Google's pre-trained word embeddings have also been incorporated to augment the feature set. The evaluation metrics, i.e. the output of our model comprises of four entities:

- Accuracy
- Senstivity
- Specificity
- Precision

For the remaining RNN based algorithms, Google word2vec embeddings [Mikolov et al. 2013a] were used as the input to the model. Our work demonstrates a comparative analysis of the results obtained by setting the dimensions of word embeddings as 50, 100 and 200. We evaluate the RNN based algorithms with the same criteria as that of the previously described algorithms, i.e. accuracy, sensitivity, specificity, and precision.

6 CONCLUSION

With a surge in the number of electronic devices connected to the internet, the propagation of fake news has increased. In this paper, we proposed a deep learning solution to the problem and have presented a comparative analysis of how standard machine learning algorithms fair against recurrent neural networks. We observed that Bi-LSTM performed better than other algorithms with reasonable performance gains as can be deduced from the values of evaluation metrics obtained.

REFERENCES

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *CoRR* abs/1412.3555 (2014). arXiv:1412.3555 http://arxiv.org/abs/1412.3555

Niall J. Conroy, Victoria L. Rubin, and Yimin Chen. 2015. Automatic Deception Detection: Methods for Finding Fake News. In *Proceedings of the 78th ASIST Annual Meeting: Information Science with Impact: Research in and for the Community (ASIST 15)*. American Society for Information Science, Silver Springs, MD, USA, Article 82, 4 pages.

William Dubay. 2004. The Principles of Readability. *CA* 92627949 (01 2004), 631–3309.

Adam E. M. Eltorai, Syed S. Naqvi, Soha Ghanian, Craig P. Eberson, Arnold-Peter C. Weiss, Christopher T. Born, and Alan H. Daniels. 2015. Readability of Invasive Procedure Consent Forms. *Clinical and Translational Science* 8, 6 (2015), 830–833. https://doi.org/10.1111/cts.12364 arXiv:https://ascpt.onlinelibrary.wiley.com/doi/pdf/10.1111/cts.12364

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (Nov. 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

S. Karmakar and Y. Zhu. 2010. Visualizing multiple text readability indexes. In *2010 International Conference on Education and Management Technology*. 133–137. https://doi.org/10.1109/ICEMT.2010.5657684

Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. http://arxiv.org/abs/1301.3781

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'13)*. Curran Associates Inc., USA, 3111–3119. http://dl.acm.org/citation.cfm?id=2999792.2999959

Veronica Perez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic Detection of Fake News. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 3391–3401. https://www.aclweb.org/anthology/C18-1287

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A Stylometric Inquiry into Hyperpartisan and Fake News. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 231–240. https://doi.org/10.18653/v1/P18-1022

Maciej Szpakowski. 2017. *Fake News Corpus.* https://github.com/several27/FakeNewsCorpus