Fake-news detection based on text classification

Nadia Sevilla; Mayelyn García

CAP4767 Data Mining· Data Analytics Program. Miami Dade College

nadia.inswasty001@gmail.com

Abstract- The increase in accessibility and availability of information on social media and internet has made it exponentially difficult to distinguish between false and true information. An important goal in improving the trustworthiness of information is being able to identify the fake news. This paper aims at investigating the performance of different classification models for detecting fake news articles, titles, and comments and evaluating the corresponding performance.

Index Terms- *Fake News Detection; Classification Model; Text Mining; Feature Extraction; Data Mining.*

I. INTRODUCTION

Fake news is false or misleading information presented as news. Fake news often has the aim of damaging the reputation of a person or entity or making money through advertising revenue. Fake news has been existing for a long time, since the "Great moon hoax" published in 1835 [1]. More recently, due to the developments and widespread usage of social networks, fake news for various commercial and political purposes have been appearing in large numbers and getting widely spread online. With deceptive words and appealing format, users can get infected by these fake news easily, which has brought about tremendous effects on the offline society; an example of this is the 2016 US president elections: According to a post-election statistical report [2], online social networks than 41.8% of the fake news data traffic in the election, which is much greater than the data traffic shares of both traditional TV/radio/print medium and online search engines respectively at that time.

Problem Studied: In this paper, we intend to study the fake news detection using diverse classification algorithms and evaluating their effectiveness. Based on various types of heterogeneous information sources, we aim at identifying fake News classifying the text parameters and finding the model which best fits the problem.

II. DATASET AND MODELS SELECTED

A. Dataset Creation

In this project there where different datasets which were merged into one to fulfill the project requirements of a binary classification dataset containing at least 200,000 observations.

The first dataset was scrapped from PolitiFact. PolitiFact website is operated by the Tampa Bay Times, where the reporters and editors can make fact-check regarding the statements made by the Congress members, White House, lobbyists, and other political groups. PolitiFact collects the political statements from the speech, news article report, online social media, etc., and will publish both the original statements, evaluation results, and the complete fact-check report at both PolitiFact website and via its official Twitter account. The statement evaluation results will clearly indicate
the credibility rating, ranging from *True* for completely accurate statements to *Pants on Fire!* for totally false claims [3].

The second dataset is the Word Embedding over Linguistic Features for Fake News Detection (WELFake) dataset which contains labeled headers and news with a label that determines if it is a real or fake article [4]. This dataset is part of an ongoing research on "Fake News Prediction on Social Media Website" as a doctoral degree program of Mr. Pawan Kumar Verma and is partially supported by the ARTICONF project funded by the European Union's Horizon 2020 research and innovation program.

The third dataset was downloaded from a Kaggle project: "Gathering real news for Oct-Dec 2016". The fourth and fifth datasets were acquired from FakeNewsNet [5].

The sixth and seventh dataset were fetched from the ISOT Fake News Dataset. articles, obtained from different legitimate news sites and sites flagged as unreliable by Politifact.com [6]. The eight dataset was taken from a public repository that focuses on the detection of COVID19-related fake news in English. The sources of data are various social-media platforms such as Twitter, Facebook, Instagram, etc. Given a social media post, the objective of the shared task is to classify it into either fake or real news [7].

Receiving Date:
Accepting Dtate:

Finally, the ninth dataset contains disinformation cases as collected by the EUvsDisinfo project. The project was started in 2015 and identifies, compiles, and exposes disinformation cases originating in pro-Kremlin media that are spread across the EU and Eastern Partnership countries. Their data base contains over 7000 disinformation cases and debunks [8]. These datasets were merged and shuffled; then the text and the label pertaining to the authenticity were exported as the only columns.

The next part of the dataset creation process was the addition of features: Principally emotion-detection; to achieve this a Neural Network model was trained to predict the proportion of each one of six emotions: Anger, sadness, fear, joy, surprise, and love; with an accuracy of 93.42%. This model was set to predict the dominant emotion present in each of the text obtained in the previous stage of the process as a categorical variable, along with the values for each emotion as numerical variables as can be seen in Picture 1.

	statement	target	anger	sadness	fear	joy	surprise	love	main_emotion
0	Central African Republic president Faustin-Arc...	True	0.528184	0.129380	0.290004	0.046111	0.002159	0.004162	anger
1	Paris Jackson addresses family issues after Ja...	True	0.282781	0.200408	0.322210	0.162250	0.014220	0.018130	fear
2	The U.S. House of Representatives ethics commi...	True	0.630080	0.106294	0.072769	0.181953	0.001379	0.007525	anger
3	EU In this October 21, 2016 photo provided by ...	False	0.385026	0.351781	0.212126	0.040994	0.001433	0.008641	anger
4	A huge rally of 10,000 Chicagoans didn t get t...	True	0.372579	0.197948	0.213057	0.208442	0.004096	0.003879	anger
...	...	...	...	...	...	...	...	...	...
203549	The views expressed herein are the views of th...	True	0.667137	0.111776	0.040695	0.165803	0.003123	0.011466	anger
203550	\nThe Liberal media just can't give Trump his ...	False	0.368812	0.231943	0.102861	0.280618	0.004170	0.011596	anger
203551	JEDDAH (Reuters) - Arab countries and Qatar sh...	False	0.624848	0.110043	0.097962	0.153164	0.001184	0.012800	anger
203552	Who says President Trump isn t moving on his a...	False	0.096386	0.058199	0.013194	0.736530	0.000972	0.094719	joy
203553	Are Native Americans Part of the Ten Lost Trib...	False	0.050304	0.491055	0.026678	0.406531	0.005482	0.019950	sadness

Picture 1. The state of the artificially created dataset.

Once this process was completed, new parameters were added using the Textstat library [9]. The added values were the following:

- Flesch Reading Ease: A formula that seeks to measure how easy it is to understand a document in English [10]; a more complete explanation of the values can be found in Table 1.

- Monosyllable rate: The proportion of monosyllabic words with respect to the quantity of total words in the text.

- Polysyllable rate: The proportion of polysyllabic words with respect to the quantity of total words in the text.

- Reading time: The estimated reading time of the given text assuming 14.69ms per character [11].

- SMOG Index: SMOG stands for 'Simple Measure of Gobbledygook'. It is a readability framework. It measures how many years of education the average person needs to have to understand a text. It is best for texts of 30 sentences or more. As an example: A score of 9.3 means that a ninth grader would be able to read the document [12].

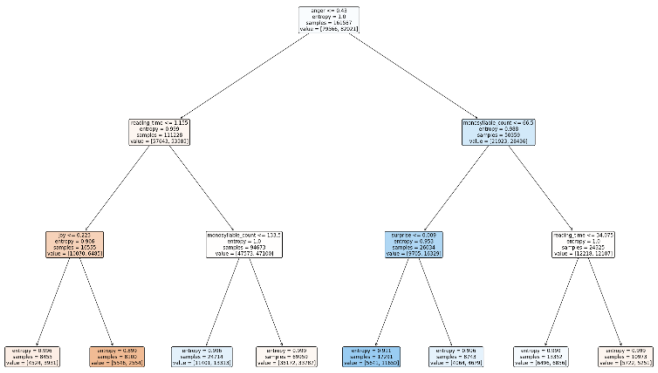
SCORE	DIFFICULTY
90 and above	Very Easy
80-89	Easy
70-79	Fairly Easy
60-69	Standard
50-59	Fairly Difficult
30-49	Difficult
29 and below	Very Confusing

Table 1. The example provided by Textstat to understand the value returned as Flesch Reading Ease.

B. Tested Models

The models employed to classify the news into real and fake were the following:

- Classification tree: A classification tree is a structural mapping of binary decisions that lead to a decision about the class of an object [13]. Although sometimes referred to as a decision tree, it is more properly a type of decision tree that leads to categorical decisions. After pre-processing the categorical variables and splitting the training data and test data in a proportion of 80/20, the results of the classification for a tree before "prunning" were as shown in Table 3. The resulting tree is the one displayed in Picture 2.



Picture 2. The classification tree obtained before pruning after fitting the training data.

	Precision	Recall	F1-score	Support
0 (Real)	0.53	0.65	0.59	19884
1 (Fake)	0.57	0.44	0.50	20513
Accuracy			0.55	40397
Macro avg	0.55	0.55	0.54	40397
Weighted avg	0.55	0.55	0.54	40397

Table 2. Classification report obtained for the pre-pruned tree.

- Naïve-Bayes: Naive Bayes is a technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable [14]. The results obtained by this classifier are as defined in Table 3.

	Precision	Recall	F1-score	Support
0 (Real)	0.53	0.23	0.32	19884
1 (Fake)	0.52	0.80	0.63	20513
Accuracy			0.52	40397
Macro avg	0.52	0.52	0.47	40397
Weighted avg	0.52	0.52	0.48	40397

Table 3. Classification report obtained for the Naïve-Bayes model.

- Logistic regression: The mathematical concept of logistic regression is to express the relationship between outcome variable and predictor variables (independent variables) in terms of logit: The natural logarithm of odds [15]. The result obtained can be seen in Table 4.

	Precision	Recall	F1-score	Support
0 (Real)	0.53	0.49	0.51	19884
1 (Fake)	0.54	0.58	0.56	20513
Accuracy			0.54	40397
Macro avg	0.54	0.54	0.53	40397
Weighted avg	0.54	0.54	0.54	40397

Table 4. Classification report obtained for the logistic regression model.

- Random Forest: Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned [16]. In this case the parameters used were 300 estimators, Gini Index as a criterion and a maximum of 4 features. The yielded results are documented in Table 5.

	Precision	Recall	F1-score	Support
0 (Real)	0.36	0.35	0.35	19884
1 (Fake)	0.38	0.39	0.38	20513
Accuracy			0.37	40397
Macro avg	0.37	0.37	0.37	40397
Weighted avg	0.37	0.37	0.37	40397

Table 5. Classification report obtained for the Random Forest model.

- Bagging: Bagging, also known as bootstrap aggregation, is a learning method that is commonly used to reduce variance within a noisy dataset [17]. The results obtained by bagging are registered on Table 6.

	Precision	Recall	F1-score	Support
0 (Real)	0.36	0.35	0.36	19884
1 (Fake)	0.38	0.38	0.38	20513
Accuracy			0.37	40397
Macro avg	0.37	0.37	0.37	40397
Weighted avg	0.37	0.37	0.37	40397

Table 6. Classification report obtained for the bagging model.

- K-nearest neighbors: A type of classification where the function is only approximated locally and all computation is deferred until function evaluation [18]. The results obtained from this method are registered on Table 7.

	Precision	Recall	F1-score	Support
0 (Real)	0.46	0.45	0.46	19884
1 (Fake)	0.48	0.48	0.48	20513
Accuracy			0.47	40397
Macro avg	0.47	0.47	0.47	40397
Weighted avg	0.47	0.47	0.47	40397

Table 7. Classification report obtained for the K-nearest neighbors classifier.

CONCLUSIONS

In this paper, six classification methods were used to perform a classificatory analysis of fake news and made predictions. We present our work and demonstrate the advantages of the data mining techniques including logistic regression and decision tree to fake news detection and classification. Taking the

1. Las notas de pie de página deberán estar en la página donde se citan. Letra Times New Roman de 8 puntos

overall model accuracy as a criterion, the models with a better performance were the classification tree and the logistic regression models, both with an accuracy of 0.55 and 0.54 respectively.

REFERENCES

Great moon hoax. https://en.wikipedia.org/wiki/Great Moon Hoax. [Online; accessed December-12-2022].

H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. Journal of Economic Perspectives, 2017.

The Principles of the Truth-O-Meter: PolitiFact's methodology for independent fact-checking https://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i. [Online; accessed December-12-2022].

Pawan Kumar Verma, Prateek Agrawal, & Radu Prodan. (2021). WELFake dataset for fake news detection in text data [Data set]. In IEEE Transactions on Computational Social Systems (0.1, Numbers doi: 10.1109/TCSS.2021.3068519, pp. 1–13). Zenodo. https://doi.org/10.5281/zenodo.4561253

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2018). Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media. arXiv preprint arXiv:1809.01286.

Ahmed H, Traore I, Saad S. (2017) "Detection of Online Fake News Using N-Gram. Analysis and Machine Learning Techniques. In: Traore I., Woungang I., Awad A. (eds) Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. ISDDC 2017. Lecture Notes in Computer Science, vol 10618. Springer, Cham (pp. 127-138)

A Heuristic-driven Ensemble Framework for COVID-19 Fake News Detection; Das, Sourya Dipta and Basak, Ayan and Dutta, Saikat; arXiv preprint arXiv:2101.03545; 2021

https://github.com/corriebar/euvsdisinfoR

https://pypi.org/project/textstat/

Blanco Pérez, A., & Gutiérrez Couto, U. (2002). Legibilidad de las páginas web sobre salud dirigidas a pacientes y lectores de la población general. Revista española de salud pública, 76, 321-331.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. Cognition, 109(2), 193-210.

Mc Laughlin, G. H. (1969). SMOG grading-a new readability formula. Journal of reading, 12(8), 639-646.

Buntine, W. (2020). Learning classification trees. In Artificial Intelligence frontiers in statistics (pp. 182-201). Chapman and Hall/CRC.

Russel, S., & Norvig, P. (1995). Practical Planning, Artificial Intelligence: A Modern Approach.

Abedin, T., Chowdhury, Z., Afzal, A. R., Yeasmin, F., & Turin, T. C. (2016). Application of binary logistic regression in clinical research. Department of Family Medicine, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada.

Ho, T. K. (1995, August). Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278-282). IEEE.

https://www.ibm.com/cloud/learn/bagging

Piryonesi, S. M., & El-Diraby, T. E. (2020). Role of data analytics in infrastructure asset management: Overcoming data size and quality problems. Journal of Transportation Engineering, Part B: Pavements, 146(2), 04020022.