IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

An Efficient Fake News Detection using Machine Learning

Achyut Pal[1], Ananya Majumder[2], Suparna Biswas[3], Antara Ghosal[4], Palasri Dhar[5], Sayan Roy Chaudhuri[6]
Students, Department of Electronics & Communication Engineering[1,2]
Faculty, Department of Electronics & Communication Engineering[3,4,5,6]
Guru Nanak Institute of Technology, Kolkata, India

Abstract: *Fake news and hoaxes have been there since before the advent of the Internet. The widely accepted definition of Internet fake news is: fictitious articles deliberately fabricated to deceive readers". Social media and news outlets publish fake news to increase readership or as part of psychological warfare. In general, the goal is profiting through clickbaits. Clickbaits lure users and entice curiosity with flashy headlinesor designs to click links to increase advertisements revenues. This exposition analyzes the prevalence of fake news in light of the advancesin communication made possible by the emergence of social networkingsites. The purpose of the work is to come up with a solution that can beutilized by users to detect and filter out sites containing false and misleading information. In this paper we have used simple and carefully selected features of the title and post to accurately identify fake posts. The experimental result shows maximum of 87.04% accuracy for logistic model.*

Keywords: Fake News, Machine Learning, Classifier

I. INTRODUCTION

The idea of fake news is not a novel concept. Notably, the idea has been in existenceeven before the emergence of the Internet as publishers used false and misleading information to further their interests. Following the advent of the web, more and more consumers began forsaking the traditional media channels used to disseminate information for online platforms. Not only does the latter alternative allow users to access a variety of publications in one sitting, but it is also more convenient and faster. The development, however, came with a redefined concept of fake news as content publishers began using what has come to be commonly referred to as a clickbait. Clickbaits are phrases that are designed to attract the attention of a user who, uponclicking on the link, is directed to a web page whose content is considerably below their expectations. Many users find clickbaits to be an irritation, and the result is thatmost of such individuals only end up spending a very short time visiting such sites. As per the report of researchers of [1], false news has major impact on the political situation of a society. This False news on the social media platforms can change opinions of peoples.

Detectin of fake news is a very big challenge, it is not an easy task [1]. People decisions and opinions are very much affected by the fake news[2]. Recently different researchers are working in fake news detection [3].

Machine learning is very much helpful in this field. Various researchers are using machine learning for the detection of fake news [3, 4, 5, 6]. It is observed that fake news is increasing day by day [7].

For content publishers, however, more clicks translate into more revenues as the commercial aspect of using online advertisements is highly contingent on web traffic.As such, despite the concerns that have been raised by readers about the use of clickbaits and the whole idea of publishing misleading information, there has been little effort on the part of content publishers to refrain from doing so.

At best, tech companies such as Google, Facebook [8,9], and Twitter have attempted to address this particular concern. However, these efforts have hardly contributed towards solving the problem as the organizations have resorted to denying the individuals associated with such sites the revenue that they would have realized fromthe increased traffic. Users, on the other hand, continue to deal with sites containingfalse information and whose involvement tends to affect the reader's ability to engage with actual news. The reason behind the involvement of firms such as Facebook in the

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

issue concerning fake news is because the emergence and subsequent development of social media platforms have served to exacerbate the problem. In particular, most of the sites that contain such information also include a sharing option that implores users to disseminate the contents of the web page further. Social networking sites allow for efficient and fast sharing of material and; thus, users can share the misleading information within a short time. In the wake of the data breach of millions of accounts by Cambridge Analytica, Facebook and other giants vowed to do more to stop the spread of fake news.

The work is concerned with identifying a solution that could be used to detect and filter out sites containing fake news for purposes of helping users to avoid being lured by clickbaits. It is imperative that such solutions are identified as they will prove to be useful to both readers and tech companies involved in the issue. To solve the above said issue we have presented a machine learning based fake news detection method.

II. PROPOSED SOLUTION

The proposed solution to the issue concerned with fake news includes the use of a tool that can identify and remove fake sites from the results provided to a user by a search engine or a social media news feed. The tool can be downloaded by the user and, subsequently, be appended to the browser or application used to receive news feeds. Once operational, the tool will use various techniques including those related to the syntactic features of a link to determine whether the same should be included as part of the search results.

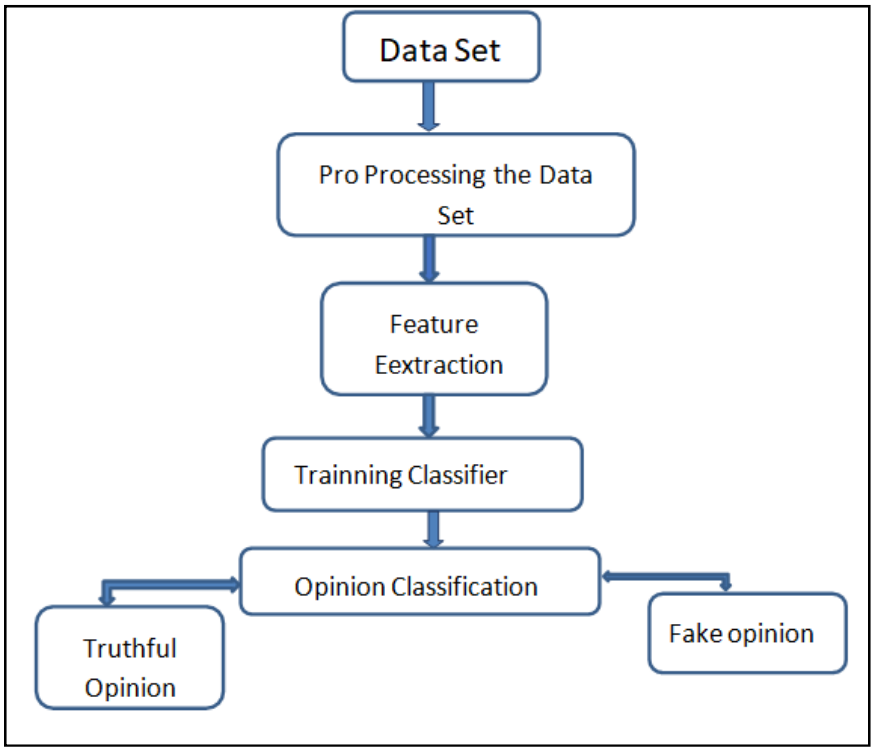

Fig.1 Flowchart of proposed technique

Sample dataset are shown in Fig.2, which contain Fake and REAL dataset. These dataset are publicly available. FAKE data is labeled with 1 and REAL data is labeled with 0.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

	URLs	Headline	Body	Label
0	http://www.bbc.com/news/world-us-canada-414191...	Four ways Bob Corker skewered Donald Trump	Image copyright Getty Images\r\nOn Sunday morn...	1
1	https://www.reuters.com/article/us-filmfestiva...	Linklater's war veteran comedy speaks to moder...	LONDON (Reuters) - "Last Flag Flying", a comed...	1
2	https://www.nytimes.com/2017/10/09/us/politics...	Trump's Fight With Corker Jeopardizes His Legi...	The feud broke into public view last week when...	1
3	https://www.reuters.com/article/us-mexico-oil-...	Egypt's Cheiron wins tie-up with Pemex for Mex...	MEXICO CITY (Reuters) - Egypt's Cheiron Holdin...	1
4	http://www.cnn.com/videos/cnnmoney/2017/10/08/...	Jason Aldean opens 'SNL' with Vegas tribute	Country singer Jason Aldean, who was performin...	1
...	...	...	...	...
4004	http://beforeitsnews.com/sports/2017/09/trends...	Trends to Watch	Trends to Watch\r\n% of readers think this sto...	0
4005	http://beforeitsnews.com/u-s-politics/2017/10/...	Trump Jr. Is Soon To Give A 30-Minute Speech F...	Trump Jr. Is Soon To Give A 30-Minute Speech F...	0
4006	https://www.activistpost.com/2017/09/ron-paul-...	Ron Paul on Trump, Anarchism & the AltRight	NaN	0
4007	https://www.reuters.com/article/us-china-pharm...	China to accept overseas trial data in bid to ...	SHANGHAI (Reuters) - China said it plans to ac...	1
4008	http://beforeitsnews.com/u-s-politics/2017/10/...	Vice President Mike Pence Leaves NFL Game Beca...	Vice President Mike Pence Leaves NFL Game Beca...	0

Fig.2 Sample dataset of FAKE & REAL

III. EXPERIMENTAL RESULT

In this experiment we have used total 8 different classifiers. All the classifiers are explained below:

3.1 Logistic Regression:

Classification is among the most important areas of machine learning, and logistic regression is one of its basic methods. Logistic regression is a fundamental classification technique. It belongs to the group of linear classifiers and is somewhat similar to polynomial and linear regression. Logistic regression is fast and relatively uncomplicated, and it's convenient for you to interpret the results. Although it's essentially a method for binary classification, it can also be applied to multiclass problems.

3.2 Decision Tree Classifier:

Decision tree is a type of supervised learning algorithm that can be used for both regression and classification problems. The algorithm uses training data to create rules that can be represented by a tree structure. Like any other tree representation, it has a root node, internal nodes, and leaf nodes. The internal node represents condition on attributes, the branches represent the results of the condition and the leaf node represents the class label. To arrive at the classification, you start at the root node at the top and work your way down to the leaf node by following the if-else style rules. The leaf node where you land up is your class label for your classification problem. Decision tree can work with both categorical and numerical data.

This is in contrast with other machine learning algorithms that cannot work with categorical data and requires encoding to numeric values. For making a decision tree, at each level we have to make a selection of the attributes to be the root node. This is known as attributes selection. This is mainly done using Gini index, Information gain and chi-square.

3.3 Random Forest Classifier:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

3.4 Stochastic Gradient Descent:

Stochastic Gradient Descent (SGD) is a simple yet very efficient approach to fitting linear classifiers and regressors under convex loss functions such as (linear) Support Vector Machines and Logistic Regression. Even though SGD has been around in the machine learning community for a long time, it has received a considerable amount of attention just recently in the context of large-scale learning. SGD has been successfully applied to large-scale and sparse machine learning problems often encountered in text classification and natural language processing. Given that the data is sparse, the classifiers in this module easily scale to problems with more than 10^5 training examples and more than 10^5 features.

3.5 Gradient Boosting Classifier

Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting. Gradient boosting models are becoming popular because of their effectiveness at classifying complex datasets, and have recently been used to win many Kaggle data science competitions. The Python machine learning library, Scikit-Learn, supports different implementations of gradient boosting classifiers, including XGBoost.

3.6 XGB Classifier

In machine learning, ensemble models perform better than individual models with high probability. An ensemble model combines different machine learning models into one. The Random Forest is a popular ensemble that takes the average of many Decision Trees via bagging. Bagging is short for “bootstrap aggregation,” meaning that samples are chosen with replacement (bootstrapping), and combined (aggregated) by taking their average. Boosting is a strong alternative to bagging. Instead of aggregating predictions, boosters turn weak learners into strong learners by focusing on where the individual models (usually Decision Trees) went wrong.

In Gradient Boosting, individual models train upon the residuals, the difference between the prediction and the actual results. Instead of aggregating trees, gradient boosted trees learn from errors during each boosting round. XGBoost is short for “extreme Gradient Boosting.” The “extreme” refers to speed enhancements such as parallel computing and cache awareness that makes XGBoost approximately 10 times faster than traditional Gradient Boosting. In addition, XGBoost includes a unique split-finding algorithm to optimize trees, along with built-in regularization that reduces overfitting. Generally speaking, XGBoost is a faster, more accurate version of Gradient Boosting. Boosting performs better than bagging on average, and Gradient Boosting is arguably the best boosting ensemble. Since XGBoost is an advanced version of Gradient Boosting, and its results are unparalleled, it’s arguably the best machine learning ensemble that we have

3.7 Multi Naïve Bayes Classifier

The multinomial Naive Bayes classifier is suitable for classification with discrete features (e.g., word counts for text classification). The multinomial distribution normally requires integer feature counts. However, in practice, fractional counts such as tf-idf may also work.

3.8 Bernoulli Naive Bayes Classifier:

BernoulliNB implements the naive Bayes training and classification algorithms for data that is distributed according to multivariate Bernoulli distributions; i.e., there may be multiple features but each one is assumed to be a binary-valued (Bernoulli, boolean) variable The Bernoulli Naive Bayes is one of the variations of the Naive Bayes algorithm in machine learning and it is very useful to use in a binary distribution where the output label may be present or absent. If you have never used this machine learning algorithm before, this article is for you. In this article, I will take you through an introduction to the Bernoulli Naive Bayes algorithm in machine learning and its implementation using Python.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

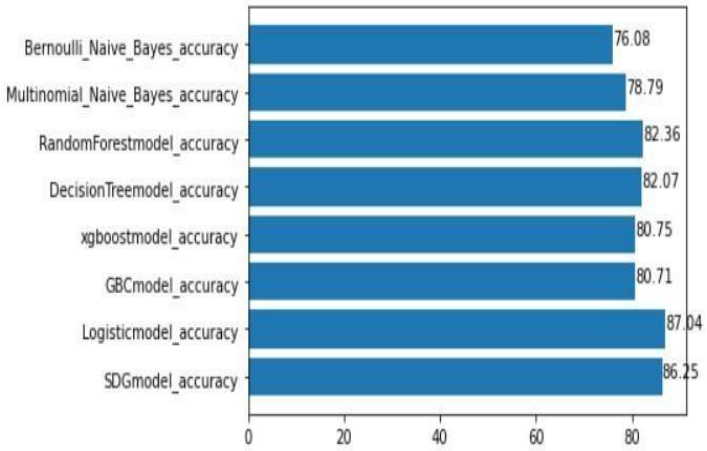

Fig.3 Experimental result of different classifiers.

IV. CONCLUSION

The task of classifying news manually requires in-depth knowledge of the domain and expertise to identify anomalies in the text. In this research, we discussed the problem of classifying fake news articles using machine learning models and ensemble techniques. The data we used in our work is collected from the World Wide Web and contains news articles from various domains to cover most of the news rather than specifically classifying political news. The primary aim of the research is to identify patterns in text that differentiate fake articles from true news. We extracted different textual features from the articles using an LIWC tool and used the feature set as an input to the models. The learning models were trained and parameter-tuned to obtain optimal accuracy. We used multiple classifiers to compare the results for each algorithm. The ensemble learners have shown an overall better score on all performance metrics as compared to the individual learners. The proposed technique provides the maximum accuracy of 87.04% for logistic model.

REFERENCES

[1] Ahmed, H., Traore, I., & Saad, S. (2017). Detection of online fake news using n-gram analysis and machine learning techniques. Proceedings of the International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, 127–138, Springer, Vancouver, Canada, 2017. https://doi.org/10.1007/978-3-319-69155-8_9.

[2] Dewey, C. (2016). Facebook has repeatedly trended fake news since firing its human editors. The Washington Post, Oct. 12, 2016.

[3] Granik, M., & Mesyura, V. (2017). Fake news detection using naive Bayes classifier. 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kiev, pp.900- 903, https://doi.org/10.1109/UKRCON.2017.810037

[4]Khan, J. Y., Khondaker, M., Islam, T., Iqbal, A., & Afroz, S. (2019). A benchmark study on machine learning methods for fake news detection. Computation and Language. https://arxiv.org/abs/1905.04749

[5]Della Vedova, M. L., Tacchini, E., Moret, S., Ballarin, G., DiPierro, M., & de Alfaro, L. (2018). Automatic online fake news detection combining content and social signals. FRUCT'22: Proceedings of the 22st Conference of Open Innovations Association FRUCT. Pages 272–279. https://dl.acm.org/doi/10.5555/3266365.3266403

[6] Wang, W. Y. (2017). "Liar, liar pants on fire": A new benchmark dataset for fake news detection. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2(Short Papers), 422–426. http://dx.doi.org/10.18653/v1/P17-2067

[7] Zhou, X., Zafarani, R., Shu, K., & Liu, H. (2019). Fake News: Fundamental theories, detection strategies and challenges. In WSDM 2019 - Proceedings of the 12th ACM International Conference on Web Search and Data Mining

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

(pp. 836-837). (WSDM 2019 - Proceedings of the 12th ACM International Conference on Web Search and Data Mining). Association for Computing Machinery, Inc. https://doi.org/10.1145/3289600.3291382

[8]Donepudi, P. K. (2020). Crowdsourced Software Testing: A Timely Opportunity. Engineering International, 8(1), 25-30. https://doi.org/10.18034/ei.v8i1.491

[9]Donepudi, P. K., Banu, M. H., Khan, W., Neogy, T. K., Asadullah, ABM., Ahmed, A. A. A. (2020b). Artificial Intelligence and Machine Learning in Treasury Management: A Systematic Literature Review. International Journal of Management, 11(11), 13-22