

Fake Media Detection based on Natural Language Processing and Blockchain Approach

Ms. T. Priyanka, M Tech student, CSE:BVCE, tpriyanka015@gmail.com

Dr. B. S. N. Murthy, Professor, CSE:BVCE, bsnmurthy2012@gmail.com

Dr. Chandra Mouli VSA, Professor CSE:BVCE, mouliac@yahoo.co.in

Gunamani Jena, Prof CSE:BVCE, drjena@gmail.com

Shubhashish Jena, CSE:OUTR sjena1998@gmail.com

Article Info

Page Number: 5385 - 5397

Publication Issue:

Vol 71 No. 4 (2022)

Abstract

Based on new technologies and developments in computer science, social media networks are an important part of people's lives. This environment has become a well-known place to share information, news, and daily reports on any topic. This is the most important time for collecting data and sending it to other places. There are some good things about this environment, but there is also a lot of fake news and information that makes it hard for people to get the right information. One of the biggest problems with this system is that there isn't enough reliable information and real news on social media. To solve this problem, we have proposed a system that combines different parts of blockchain and natural language processing (NLP) to use machine learning techniques to spot fake news and better predict fake user accounts and posts. This is done with the Reinforcement Learning method.

process. In this paper, the author uses the Deep Learning Reinforcement Algorithm to predict fake media, and it shows less MAE and RMSE than other algorithms. The MAE or RMSE is the difference between the actual values and the values that were predicted. The lower the MAE, the better the algorithm works. So, to lower MAE and RMSE values even more, we are extending this project with the XGBOOST algorithm, which is an advanced algorithm in the field of machine learning. This algorithm gives less MAE or RMSE than the proposed algorithm.

Natural Language Processing, the blockchain, fake news, and reinforcement learning are some examples of keywords.

Article History

Article Received: 25 March 2022

Revised: 30 April 2022

Accepted: 15 June 2022

Publication: 19 August 2022

INTRODUCTION:

The real part of social media is the variety of information that is shared. Since 2017, fake news has become a very important topic that 365% of people talk about online [1]. Fake news becomes an unsolved problem in the data and information consumption application layer of social networks. It also becomes a serious and difficult problem in the advancement of information that shows up in the diplomatic, economic, and political sectors. The fake information leak shows that the network

resources are being used for things that aren't necessary. It also has the totality and validity of the content based on the services that are available. [2]. So, spreading false information has an effect on the Quality of Trust (QoT), which is used to distribute news [3]. Machine learning improves the level of security needed for daily networking on social media sites. A survey by a non-government organisation found that there are a lot of fake accounts and information circulating on social networking sites. In this case, the bad and unwanted accounts need to leave the network so that the data centre has more room and the mess and political problems in the network can be handled.

Propaganda, which is only used for political purposes, is another area related to information gathering. [4]–[6]. The language used to make fake news is very clever because it is designed to make people angry and upset in order to spread false information [7–9]. Fake news detection is the ability to analyse the information based on how true it is [10, 12]. With more noisy and unstructured data, more users, and more news, there is a need for an automatic way to find fake news [13–15]. Because of recent changes in machine learning, deep learning, and artificial intelligence, these terms no longer cover everything. Proofing the authorship of digital content is one of the steps that must be taken before information can be shared.

1.1 NLP method for processing text:

To process all of the news, the author is using NLP techniques to get rid of stop words, stemming, and lemmatization, and then the TFIDF (term frequency inverse frequency document) algorithm to turn all of the words into numbers. TFIDF will replace each word with how often it is used on average.

1.2 Reinforcement Learning: This algorithm is made up of three parts: the STATE, the AGENT, and the PUNISHMENT or REWARD. The state will look at how similar the user's news is to fake or real news, and based on that similarity, the state's agent will decide if the news is fake or real. If the prediction is right, the algorithm will get a reward, but if it's wrong, a penalty will be applied. The lower RMSE will be, the higher the reward.

Blockchain technology is a decentralised network architecture where data is stored on multiple nodes/servers. This is different from centralised servers, where data is stored on a single server and anyone can access it and change the data. With Blockchain, data is stored on multiple nodes, and each piece of data is stored as a block/transaction. Each block also has a unique hash code. Before storing new data, Blockchain will check the hash code of the previous blocks. If the data hasn't changed, the same hash code will be made, and verification will be successful. If the data has changed, verification will fail, and the change will be found. Blockchain verification is called "Proof of Work," and it's called "immutable" because its data can't be changed.

1.3 Motivation: These days, almost everyone relies on online content for entertainment, news, shopping, and many other things. Some bad users may take advantage of this by putting out fake online content, which could hurt online services.

1.4 Problem Statement: With the recent growth of technology and the use of apps in everyday life, people are posting and sharing things that don't make sense on social media, which makes a mess on different social platforms. This process makes it hard to find the right answer to the question that was asked. Twitter is one of the social networks mentioned. It has a huge number of users who share millions of tweets every day about all kinds of things. In this process, machine learning and the blockchain system play a big role in stopping the spread of fake news.

1.5 The main goal of this article is to figure out if online news is fake or real, and then use Blockchain technology to make sure that the news publisher is an authorised user.

1.6 Work Scope: Social media online plays a big role in real-world events like natural disasters, elections, social movements, etc. Since more people are using social media, the amount of fake news has grown. People often use social media to spread false information by changing real news or making up new news. From a national security point of view, making and spreading fake news is dangerous in a number of ways. Hence Getting better at spotting fake news is an important step toward making online social network information more reliable. Researchers have used many different methods, algorithms, tools, and techniques over time to figure out how to spot fake news on social networks.

1.8 Order of the Report:

Five chapters make up the rest of the report. After this first chapter, chapter 2 talks about a survey of the system that is already in place. This gives an overview of the research that has been done so far in the field of predicting whether online news is fake or true, using NLP (natural language processing) to extract keywords, stop word removal, and stemming technique to preprocess news data, Reinforcement learning algorithm to predict whether online news is fake or true, and Blockchain technology to confirm that the news publisher is an authorised user. For this project, the author used the BUZZFACE dataset, which has both real and fake news. We will use this dataset to train existing Random Forest Machine Learning algorithm and propose Reinforcement learning algorithm and compare their performance in terms of RMSE (root mean square error, or the percentage of wrong predictions) and MAE.

In Chapter 3, the proposed system is explained. This begins with an explanation of the dataset and the models used in the report. Then it talks about how the proposed system is put together. Describes the process and algorithms that were used, as well as the details of the software that was used for the research. It also talks about the criteria that were used to evaluate this study.

Chapter 4 tells about the experiment and how it turned out. Smart healthcare systems can use this model to find the right way to diagnose diseases.

In Chapter 5, the results of all the models in this research paper are summed up, and suggestions are made about when to use each model. It shows how work should be done in the future.

3.1.2 XGBoost:

XGBoost is a library for optimised distributed gradient boosting that is very fast, flexible, and easy to move around. The method is used to solve supervised learning problems, and data scientists have used it a lot to get the best results for a wide range of machine learning problems.

3.1.3 Learning from rewards:

This algorithm is made up of three parts: STATE, AGENT, and PUNISHMENT or REWARD. The state will look at how similar the user's news is to fake or real news, and based on that similarity, the state's agent will decide if the news is fake or real. If the prediction is right, the algorithm will get a reward, but if it's wrong, a penalty will be applied. The lower RMSE will be, the higher the reward.

3.1.4 NLP The NLP method for processing text: To process all of the news, the author is using NLP techniques to get rid of stop words, stemming, and lemmatization, and then the TFIDF (term frequency inverse frequency document) algorithm to turn all of the words into numbers. TFIDF will replace each word with how often it is used on average.

3.1.5 Blockchain technology: Blockchain is a decentralised network architecture where data is stored on multiple nodes/servers. This is different from existing centralised servers, where data is stored on a single server and anyone can access it and change the data. In Blockchain, data is stored on multiple nodes, and each piece of data is stored as a block/transaction. Each block also has a unique hash code. Before storing new data, Blockchain will check the hash code of the previous blocks. If the data hasn't changed, the same hash code will be made, and verification will be successful. If the data has changed, verification will fail, and the change will be found. Blockchain verification is called "Proof of Work," and it's called "immutable" because its data can't be changed.

3.3 Architecture/Framework:

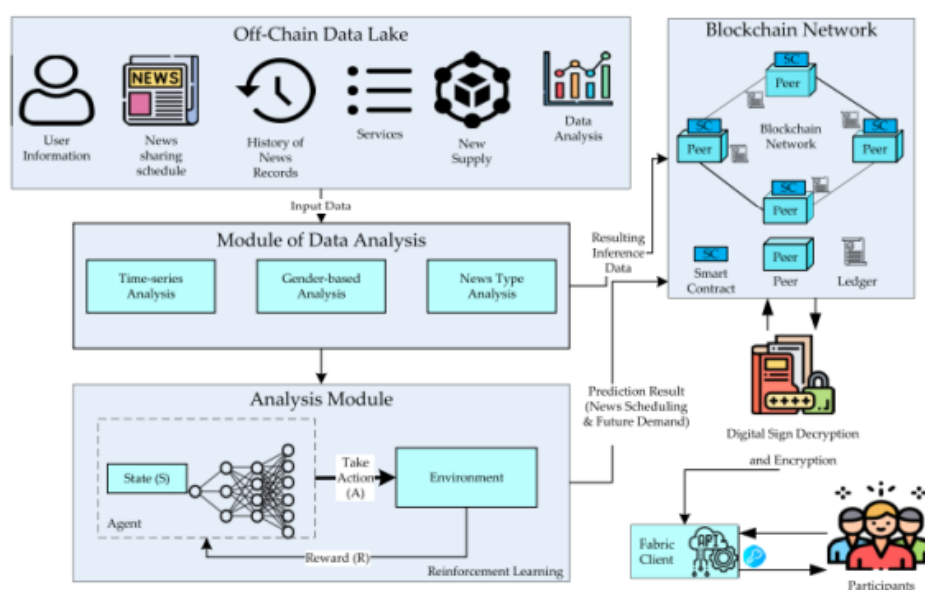


Fig 1:Architecture/Framework

3.4 Algorithm and Process Design:

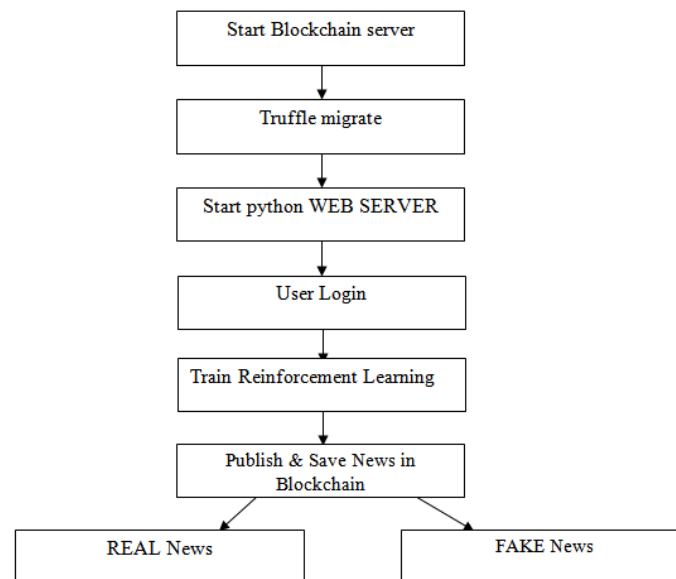


Fig 2: Algorithm and process design

1) User Signup: With this module, users can sign up for the app, and their information will be saved in Blockchain so that users can be checked.

1) User Login: This module allows users to sign in to an application

2) Train: This module lets users teach algorithms to tell the difference between fake and real news.

3) Publish News: With this module, a user can post news, and then trained Reinforcement algorithms will figure out if the news is real or fake.

4) View News: With this module, users can look at news from all over the world and tell if it's real or fake. This way, everyone can know which news is real and which isn't.

4 How It Was Done and What Happened

4.1 Getting information

The <https://github.com/gsantia/BuzzFace> dataset is the one that researchers use the most.

Below is a picture of the BUZZFACE dataset file, which can be found in the 'Dataset' folder.

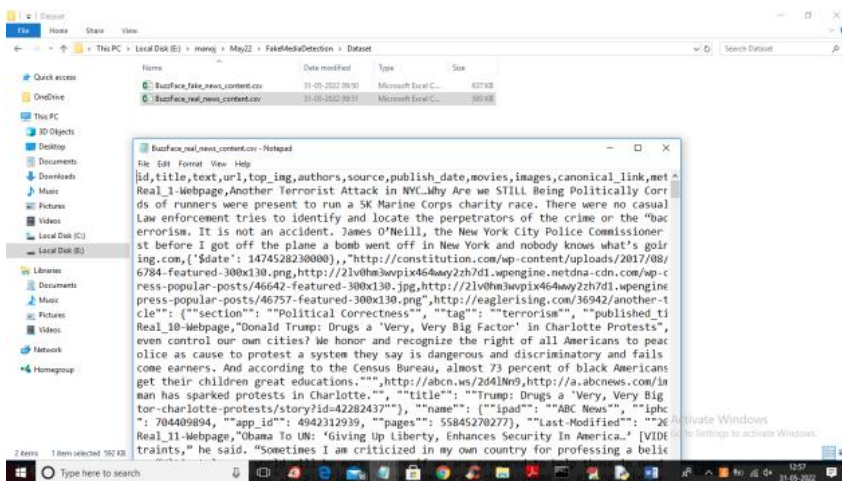


Fig 3:BUZZFACE dataset

In the picture above, there are two files, one with fake news and the other with real news. You can see the dataset in NOTEPAD as well. We use the above dataset to train algorithms.

4.2 Metrics for Evaluating:

There are different ways to measure how well machine learning techniques work. We used the mean absolute error (MAE), the root mean square error (RMSE), the mean absolute percentage error (MAPE), and the R² score in this process. The following equations were used to figure out how well the process worked.

Mean Absolute Error (MAE): The MAE is used to find the difference between the actual value and the predicted value when measuring the prediction model's performance. Equation 5 shows how a grade is given.

$$MAE = \frac{\sum_{n=1}^x |Z_1 - \hat{Z}_i|}{x}$$

Root Mean Square Error (RMSE): presents the MAE square error. This process is used for the overall evaluation of prediction model performance. RMSE range starts from 0 to ∞ . Equation 6 presents the evaluation process.

$$RMSE = \sqrt{\frac{\sum_{n=1}^x |Z_1 - \hat{Z}_i|^2}{x}}$$

R² Score: To show the variance proportion between two variables, this statistical measurement was applied. R² score is between 0 to 1.

$$R^2 Score = \frac{\sum_{n=1}^x |Z_1 - \hat{Z}_i|^2}{\sum_{n=1}^x |Z_1 - \hat{Z}_i|^2}$$

4.3 Outcome:

To run project double click on 'runServer.bat' file to start python WEB SERVER and get below figure

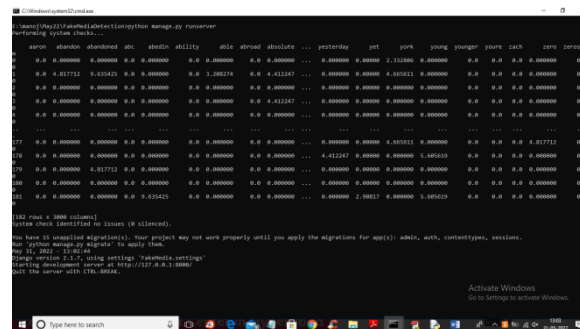


Fig 4: Python WEB SERVER

In above figure server read all dataset and then Preprocess and then convert dataset words into numeric vector and in above vector first row contains dataset word and remaining rows contain average frequency of the word. Now open browser and enter URL as <http://127.0.0.1:8000/index.html> and press enter key to get below output

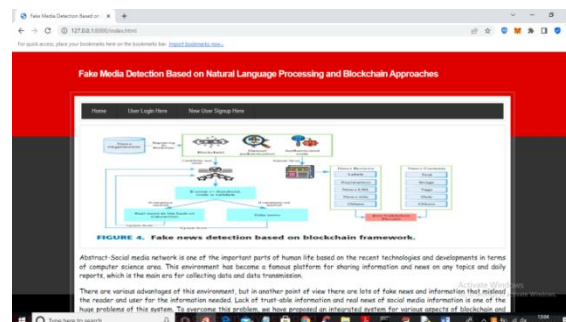


Fig 5:New user Signup here

In above figure click on 'New User Signup Here' link to get below figure



Fig 6:signing up

In above figure user is signing up and press button to get below output



Fig 7:User Login

In above figure in red colour text we can see signup process completed and now click on 'User Login' link to login as user

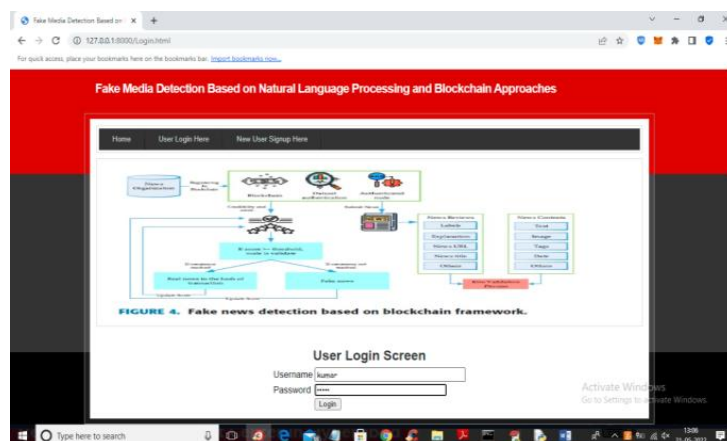


Fig 8:Login and After login

In above figure user is login and after login will get below figure

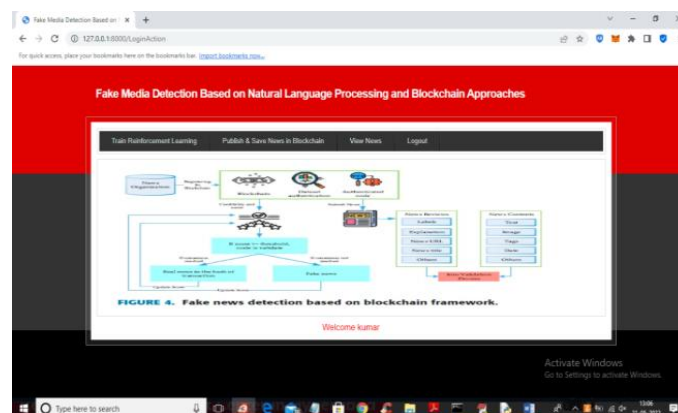


Fig 9:Train Reinforcement Learning

In above figure user can click on 'Train Reinforcement Learning' link to train algorithm and get below output



Fig 10:RMSE

In above figure existing Random forest and propose Reinforcement and with propose algorithm we got less RMSE and now click on 'Publish & Save News in Blockchain' link to publish news

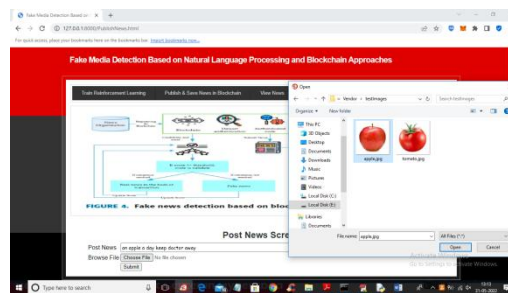


Fig 11:NEWS about Fake or True

In above figure user entered some NEWS and uploading related picture and then press button to allow application to check news as FAKE or true and then save in Blockchain

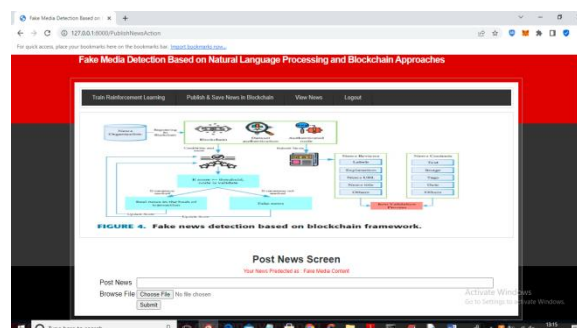


Fig 12:Showing the news is Fake

Above figure news is detected as fake and now try another news

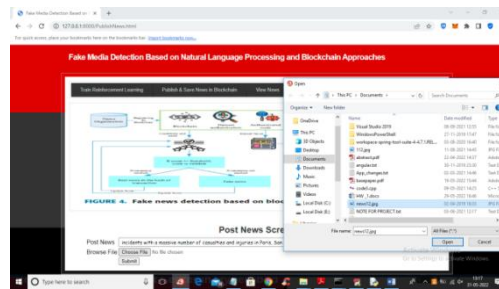


Fig 13: Entering other News to Check

In above figure entered some other news and press button to get below output

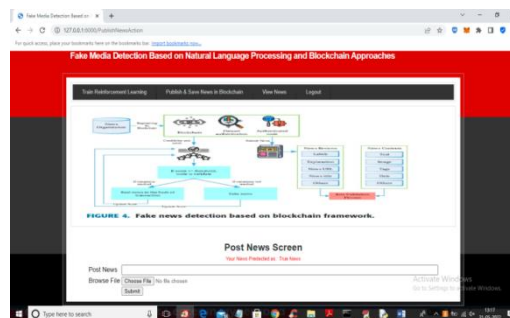


Fig 14: Showing the news is detected as True News

In above figure news is detected as 'True News'. Propose algorithm matched words from REAL and FAKE dataset and then whatever has more similarity that will be predicted. Now click on 'View News' link to get below figure

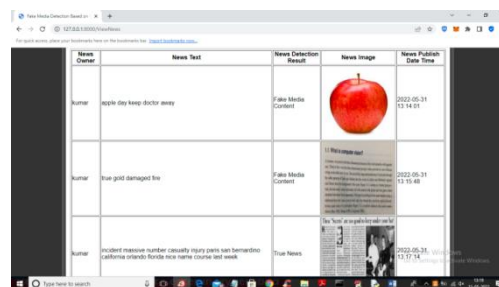


Fig 15: view news

In above figure all users can view all news with detection and prevention result as 'true or fake' and if news is fake then user will ignore it.

Extension outcomes:

In this paper author is using Deep Learning Reinforcement Algorithm to predict Fake Media and its showing less MAE (mean absolute error) and RMSE (root mean square error) compare to existing algorithms. MAE or RMSE is the error difference between actual and predicted values so the lower the MAE the better is the algorithm performance.

So to further reduce MAE and RMSE values we are extending this project with XGBOOST algorithm which is advance algorithm in machine learning field and this algorithm is giving less MAE or RMSE compare to propose algorithm.

XGBoost is an optimised distributed gradient boosting library, which is highly efficient, flexible and portable. The method is used for supervised learning problems and has been widely applied by data scientists to get optimised results for various machine learning challenges.

In below figure we have trained all existing, propose and extension algorithm with Fake Media dataset and extension algorithm is giving less MAE value

Run all modules like previous project figure shots and when we click on 'Train Reinforcement Learning' then will get

below output

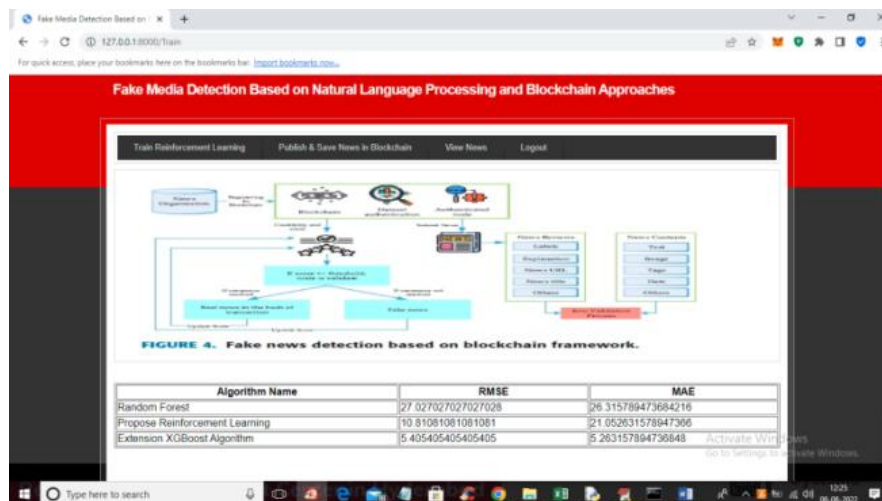


Fig 16: XGBOOST is giving less RMSE & MAE

In above figure we can see in all 3 algorithms extension XGBOOST is giving less RMSE and MAE

CONCLUSION

One of the most-studied problems in technology right now is the spread of fake news. This is because people don't feel safe or trust the news they see on social media. In this article, we talked about how blockchain and machine learning can be used together to solve problems and build a trust-based architecture for online news sharing. We used the reinforcement learning technique, which is a learning-based algorithm, to make a strong decision-making architecture and combined it with a blockchain framework, smart contracts, and a customised consensus algorithm that works well for the Proof-of-Authority protocol. Social media is an important part of this process. The platform for sharing information has fake news, and it would be good to improve and learn more about the Proof-of-Authority protocol and user validation.

Bibliography

- [1] V. P. Miletskiy, D. N. Cherezov, and E. V. Strogetskaia, “Transformations of professional political communications in the digital society (by the example of the fake news communication strategy),” in *Proc. Commun. Strategies Digit. Soc. Workshop (ComSDS)*, 2019, pp. 121–124.
- [2] N. R. de Oliveira, D. S. V. Medeiros, and D. M. F. Mattos, “A sensitive stylistic approach to identify fake news on social networking,” *IEEE Signal Process. Lett.*, vol. 27, pp. 1250–1254, 2020.
- [3] G. Liu, Y. Wang, and M. Orgun, “Optimal social trust path selection in complex social networks,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 24, 2010, pp. 1391–1398.
- [4] M. N. Nikiforos, S. Vergis, A. Styliadou, N. Augoustis, K. L. Kermanidis, and M. Maragoudakis, “Fake news detection regarding the Hong Kong events from tweets,” in *Proc. Int. Conf. Artif. Intell. Appl. Innov. Greece: Springer*, 2020, pp. 177–186.
- [5] A. R. Merryton and G. Augasta, “A survey on recent advances in machine learning techniques for fake news detection,” *Test Eng. Manag.*, vol. 83, pp. 11572–11582, 2020.
- [6] X. Dong, U. Victor, S. Chowdhury, and L. Qian, “Deep two-path semisupervised learning for fake news detection,” 2019, arXiv:1906.05659. [Online]. Available: <http://arxiv.org/abs/1906.05659>
- [7] S. Kumar, R. Asthana, S. Upadhyay, N. Upreti, and M. Akbar, “Fake news detection using deep learning models: A novel approach,” *Trans. Emerg. Telecommun. Technol.*, vol. 31, no. 2, p. e3767, Feb. 2020.
- [8] P. Bahad, P. Saxena, and R. Kamal, “Fake news detection using bidirectional LSTM-recurrent neural network,” *Procedia Comput. Sci.*, vol. 165, pp. 74–82, Jan. 2019.
- [9] G. Sansonetti, F. Gasparetti, G. D’Aniello, and A. Micarelli, “Unreliable users detection in social media: Deep learning techniques for automatic detection,” *IEEE Access*, vol. 8, pp. 213154–213167, 2020.
- [10] M. Mahyoob, J. Algaraady, and M. Alrahaili, “Linguistic-based detection of fake news in social media,” *Int. J. English Linguistics*, vol. 11, no. 1, p. 99, Nov. 2020.
- [11] A. Koirala, “COVID-19 fake news classification using deep learning,” *Tech. Rep.*, 2020. [Online]. Available: <https://www.cs.ait.ac.th/xmlui/handle/123456789/981>
- [12] H. Gill and H. Rojas, “Chatting in a mobile chamber: Effects of instant messenger use on tolerance toward political misinformation among south Koreans,” *Asian J. Commun.*, vol. 30, no. 6, pp. 470–493, Nov. 2020.

- [13] J. L. Alves, L. Weitzel, P. Quaresma, C. E. Cardoso, and L. Cunha, “Brazilian presidential elections in the era of misinformation: A machine learning approach to analyse fake news,” in Proc. Iberoamerican Congr. Pattern Recognit. Havana, Cuba: Springer, 2019, pp. 72–84.
- [14] N. R. de Oliveira, P. S. Pisa, M. A. Lopez, D. S. V. de Medeiros, and D. M. F. Mattos, “Identifying fake news on social networks based on natural language processing: Trends and challenges,” *Information*, vol. 12, no. 1, p. 38, Jan. 2021.
- [15] D. Mouratidis, M. N. Nikiforos, and K. L. Kermanidis, “Deep learning for fake news detection in a pairwise textual input schema,” *Computation*, vol. 9, no. 2, p. 20, Feb. 2021.
- [16] A. Narayanan, J. Bonneau, E. Felten, A. Miller, and S. Goldfeder, *Bitcoin Cryptocurrency Technologies: A Comprehensive Introduction*. Princeton, NJ, USA: Princeton Univ. Press, 2016.
- [17] R. Raturi, “Machine learning implementation for business development in real time sector,” *Int. J. Pure Appl. Math.*, vol. 119, no. 15, pp. 1289–1300, 2018.
- [18] B. Liu, Q. Zhao, Y. Jin, J. Shen, and C. Li, “Application of combined model of stepwise regression analysis and artificial neural network in data calibration of miniature air quality detector,” *Sci. Rep.*, vol. 11, no. 1, Dec. 2021, Art. no. 3247.
- [19] A. Karbowski, “A note on patents and leniency,” *Gospodarka Narodowa*, vol. 301, no. 1, pp. 97–108, 2020.
- [20] J. A. Vijay, H. A. Basha, and J. A. Nehru, “A dynamic approach for detecting the fake news using random forest classifier and NLP,” in *Computational Methods and Data Engineering*. Springer, 2021, pp. 331–341.