Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Design and Detection of fake News in Social Plaforms using Machine Learning

Ms. Supriya Ashok Bhosale[1], Dr. Lokendra Singh Songare[2]

[1]Ph.D. Scholar, Department of Computer Science & Engineering, Dr. A. P. J. Abdul Kalam University, Indore, MP, India

[2]Assistant Professor, Department of Computer Science & Engineering, Dr. A. P. J. Abdul Kalam University, Indore, MP, India

Article Info

Page Number: 9784 - 9796

Publication Issue:

Vol 71 No. 4 (2022)

Article History

Article Received: *15 September 2022*

Revised: *25 October 2022*

Accepted: *14 November 2022*

Publication: *21 December 2022*

Abstract

The overview of the internet and the quick adoption of public news platforms (such as Facebook(FB), Twitter and Instagram) prepared the door for unprecedented levels of knowledge distribution in human history. Thanks to social media platforms, consumers are creating and sharing more knowledge compared to before, Most of it is incorrect and has no bearing on the discussion. It's difficult to categorise a written work as misleading or disinformation using an algorithm. Even an expert in a given field must consider a variety of factors before deciding whether or not an item is true. For detecting spurious news, researchers recommend using a machine learning classification approach. Our research looks into different textual qualities that can be used to tell the difference between false and real content. We train a set of distinct machine learning algorithms using diverse integral approaches and evaluate their performance on real-world datasets using those properties. Our proposed ensemble learner method outperforms individual learners.

Keywords: Fake News, Machine Learning ,Deep Learning,

Introduction

People today often get their news from social media sites like Facebook, WhatsApp, Twitter, and Telegram. They trust this information without first checking to see if it's true or figuring out where it came from. People from all over the world are getting more and more likely to spread false information because social media platforms make it easy, cheap, and accessible to share

information. It is possible to make money or get other benefits by giving the public false information on purpose. It can be used to hurt important people, change government policy, and do other things that help oneself. Because of this, many different research projects have been done to find fake news with a high degree of accuracy and stop the bad things that can happen when it spreads. As a response to the worries mentioned in the last sentence, this paper gives an in-depth look at the ways that can now be used to spot fake news. In recent years, collecting and analysing huge amounts of information has become much easier thanks to the Internet and other social media. We spend most of our waking hours on social media, where it's easy to share and talk about news with friends and other readers. Because of this, we tend to keep an eye on or get our news from web-based platforms rather than traditional news organisations. This is because traditional news sources don't make it easy to share and talk about the news. Even though social media has a lot of benefits, the quality and standard of the stories posted on these platforms is much lower than that of traditional news sources. The rise of different social media platforms has helped the media industry as a whole because it has made it easier for publications to give their readers timely and up-to-date news, which has helped the media sector. Articles in online news sources that contain information that has been made up are put out with the goal of swaying public opinion in order to make money or get power. Some people are worried that if fake news keeps getting around, it could hurt both people and groups. The spread of fake news could throw off the delicate balance of trustworthiness that exists in the news ecosystem. People's ideas about the world will always change in some way as a direct result of spreading false information. Most of the time, spammers use clickbait and fake news stories to make money from online advertising. Fake news is one of the biggest problems for businesses, journalists, and democracies all over the world, and it has terrible effects in every part of the world. When the news came out that U.S. President Barack Obama had been hurt in an explosion, the stock market fell by $130 billion. The sudden shortage of salt in Chinese supermarkets because of a false report that iodized salt would help counteract the effects of radiation after the Fukushima nuclear leak in Japan [28] is another example of how much damage fake news campaigns can do. Another example is how things got worse between India and Pakistan after the Bagalkot strike was wrongly reported. This led to the deaths of military personnel and the loss of expensive military equipment. Most of the people who use social media don't think that it has such a big impact on society. In the best case, this will cause data articles to be made that are not possible at all. Fake news websites put out intentionally false information that they know isn't

true, but they still promote it. The Internet runs on how people interact with each other. The main goal of websites like these is to spread false information in order to change people's minds.

Related work

Gupta et al.,[1]Using the fake news combat spectrum, we look at how different groups have fought back against the spread of false information. In the last part of this article, we'll talk about how technology can help us deal with the problems and opportunities that fake news brings.

D. Rohera et al[2]People today often get their news from social media sites like Facebook, WhatsApp, Twitter, and Telegram. They trust this information without first checking to see if it's true or figuring out where it came from. People from all over the world are getting more and more likely to spread false information because social media platforms make it easy, cheap, and accessible to share information. It is possible to make money or get other benefits by giving the public false information on purpose. It can be used to hurt important people, change government policy, and do other things that help oneself. Because of this, many different research projects have been done to find fake news with a high degree of accuracy and stop the bad things that can happen when it spreads. Because of all the worries.

E. Elsaeed, et al[3]This work describes a system for finding fake news that is made up of a set of voting classifiers and algorithms for extracting features and choosing features. The suggested method can tell the difference between fake and true reports of events. We started by cleaning up the data by getting rid of duplicates and wrong characters and setting a maximum size for the dictionary (lemmatization). Second, we used two different feature extraction procedures to get information that was relevant to the problem. They were the term frequency-inverse document frequency methodology and the document to vector algorithm, which is a word embedding strategy.

H. Saleh, et al.[4]The main goal of this investigation is to find the model that is best at predicting performance with high levels of accuracy. Because of this, we recommend using a model of convolutional neural networks that is perfect for spotting false information (OPCNN-FAKE). We use four different fake news benchmark datasets to compare the OPCNN-FAKE to Recurrent Neural Networks (RNNs), Long Short-Term Memories (LSTMs), and the six traditional Machine Learning approaches (Decision Trees [DTs], Logistic Regression [LRs], K-Nearest Neighbors [KNNs], Random Forests [RFs], Support Vector Machines [SVMs], and Naive Bayes [NBs]).

M. Narra et al.,[5] In this study, the accuracy of the models used to find hoaxes will be measured so that the effects of using different subset feature selection procedures can be compared. With the help of machine learning and pre-trained deep learning models, Principal Component Analysis and Chi-Square are looked into to help choose features. A number of different preprocessing steps are also looked into to see how they affect the ability to spot fake news.

M. Umer et al.[6]In this body of work, using dimensionality reduction techniques was suggested as a way to give the classifier easier-to-manipulate feature vectors. For this study, a dataset from the Fake News Challenges (FNC) website was used to figure out how the logic worked. This set of data had four categories: agree, disagree, discuss, and not related. People look to nonlinear properties for help in a lot of situations.

Proposed methodology

To properly deal with the problem of spotting fake news on social media, it is clear that a practical solution must include a number of different parts. The Naive Bayes classifier, Support Vector Machines, and semantic analysis all work together to form the foundation of the proposed method. Instead of using algorithms that can't mimic cognitive functions, the proposed method is made up of only Artificial Intelligence techniques, which are necessary for separating real information from fake information. The third part of the strategy is the combination of natural language processing (NLP) techniques and machine learning algorithms, both of which can be further broken down into unsupervised learning strategies. Even though each of these methods can be used on its own to spot and identify fake news, they have been combined into a single algorithm to improve accuracy and make the algorithm suitable for use on social media. This was done to make the way fake news is found better. And because SVM and Naive Bayes classifier are both good supervised learning algorithms, they often "compete" with each other when it comes to the classification problem. Trials have shown that neither the Support Vector Machine nor the Naive Bayes classifier are very good at spotting fake news. So, the proposed method puts a lot of emphasis on combining the two in order to get a higher level of accuracy. The authors of the paper "Combining Naive Bayesian and Support Vector Machine for Intrusion Detection System" combine the two methods to improve the accuracy of classification done by either the SVM or the Naive Bayes classifier. Their research showed that their "hybrid algorithm" greatly reduced "false positives" and increased the number of times a balance was found. In terms of accuracy, it was better than both the SVM and the Nave Bayes classifier. Even though Intrusion Detection Systems (IDS) were used in this test, it's easy to

see how the same idea could be used to spot fake news. One way to improve the performance of an algorithm is to add semantic analysis to support vector machines and naive bayes classifiers. Using the information from the last two methods, this can be done. The main problem with the Naive Bayes classifier is that it thinks that all of the features in a document (or other textual format) are independent, which is almost never the case in real life. When everything is handled as if it were unrelated to everything else, accuracy goes down and hard-to-understand connections aren't found. As we have seen, one of the most important benefits of this type of research is that it can find connections between different words. One of the biggest problems with the Naive Bayes classifier is that it doesn't take into account the meaning of words. One way to improve the performance of a classifier is to use both semantic analysis and support vector machines (SVM). The phrase "focused attention of Support Vector Machines into informative subspaces of the feature spaces" is used by the author of the paper "Support Vector Machines for Text Categorization Based on Latent Semantic Indexing" to show how the two methods can be used together to improve productivity. In the context of the experiment, semantic analysis was able to capture the "underlying substance of material in semantic meaning." As a result, SVM became more efficient because it could spend less time classifying input that didn't mean anything and more time doing semantic analysis to organise useful data. As was just said, semantic analysis's main benefit is that it can use the relationships between words to find important data that can be used to improve SVM. Semantic mining can be used to do this.

The Support Vector Machine, which is sometimes called SVM, is a popular choice among supervised machine learning algorithms for both classification and regression work. Even though we can use it for regression analysis, its best use is for classification, so that's where we'll focus. The goal of the Support Vector Machine (SVM) technique is to find a hyperplane in an N-dimensional space that clearly splits the data into classes. The number of attributes is what determines the size of the hyperspace. If there are only two features that are used to make the hyperplane, the result will be a straight line. When there are only three input features, the hyperplane breaks down into a two-dimensional plane. If something has more than three things that make it different from other things, it can be hard to picture everything.

Let's say that x1 and x2 are the independent variables, and the colour of the circle, blue or red, is the dependent variable.

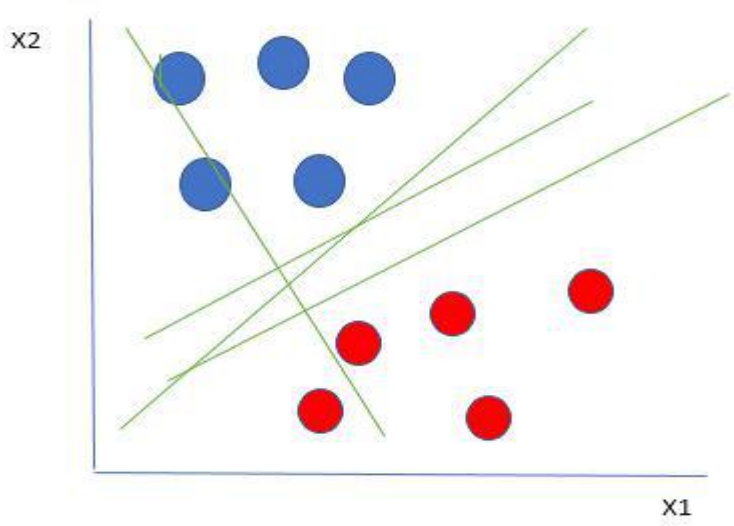

Figure 1-SVM parameters with different hyper plane

The diagram shows that our data can be split up along different lines, which is the case (our hyperplane here is a line because we are examining only two input features x1, x2). Next, we need to figure out how to choose the best line, or more generally, the hyperplane that will divide our data the best.

The SVM's "heart": The SVM kernel is a function that makes an input space with fewer dimensions into one with more. This turns a problem that can't be solved separately into a problem that can be solved separately. It works best when it is used to solve nonlinear separation problems. Before deciding how to split the data based on the output labels, the kernel goes through a series of incredibly complex data transformations. Before deciding how to split the data, this is what the kernel does at its most basic level.

Steps for PCA algorithm

Taking the set of data into account in the analysis

In the first step, we take the dataset we were given and split it into a training set (X) and a validation set (Y) (Y).

Representing data into a structure

The next step is to make a picture of the data set we have. The two-dimensional matrix will be used to show X, which is our independent variable. In this table, a row stands for a data item and a

column stands for a feature. The dimension that matches the dimensions of the data collection is the total number of columns.

Standardizing the data

Our data set will be standardised at this point. For example, features with a higher variance in a certain column are better than those with a lower variance in that column.

To test this theory, we will first divide each data point in a column by the column's standard deviation and then compare the two sets of results. Let's call the matrix Z in this particular case.

The Covariance Analysis for Z

To find the covariance of a matrix Z, the matrix is turned upside down. The next step is to multiply the value by Z after we have changed it. The covariance matrix for Z will be what the matrix gives back.

Obtaining Eigen Values and Eigen Vectors

Right now, you need to find the Z-covariance matrix's eigenvalues and eigenvectors. The most information is in the directions that lie along the axes of the covariance matrix or the eigenvectors. Also, the coefficients of these eigenvectors are called eigenvalues.

Changing how the Eigen Vectors are arranged

In this step, we take all of the eigenvalues and put them in "decreasing order," which just means that we arrange them from biggest to smallest. Also, change the order of the eigenvectors in the eigenvalue matrix P so that they are in the right place. The matrix P* will be used to describe the end result.

On the other hand, new features are calculated and major factors are taken into account.

During this step, the properties that have just been calculated will be looked at. In order to reach this goal, we will multiply the P* matrix by Z. Each item in the Z* matrix that was made shows how the initial observations were put together in a linear way. The columns of the Z* matrix can be thought of as separate things.

Take out any parts of the newly created data set that don't add anything interesting or important.

Since a new set of skills has been developed, we can now choose which ones to keep and which ones to get rid of. This means that only the most important features will be added to the new dataset, while the less important ones will be taken out.

Results Analysis

The suggested ensemble model of SVM Kernel is then trained on 49,972 samples, and in the final round of tests, it is tested on 25,413 headlines and articles. To finish the course, we use a Dell PowerEdge T430 with 32 gigabytes of DDR4 RAM, a graphics processing unit with 2 gigabytes of memory, and the course (RAM). Training time is figured out by running epochs on the "Fake News Challenge Dataset" with word embeddings that have already been trained and showing the classification results. The whole thing takes about three hours. On the other hand, it takes 1.8 hours to figure out how to reduce the number of features. An SVM Kernel design with fewer features, principal component analysis, and chi-square analysis were looked into, and the results were looked at. Based on the study of the results, it was decided that principal component analysis (PCA) is better than other methods for reducing the number of dimensions when things are bad. This is because it is more accurate. The accuracy of the model shown here is 97.8%, which is higher than the accuracy of other models. As you can see in, the average F1 score, recall score, and precision score were all in the range of 97.4%, 98.2%, and 97.8%, respectively, for all classes. In, you can see all of the statistical results that we got from our method. Based on how important the data is from a statistical point of view, our method can be used to figure out if a piece of news is real or not. what topics were covered in the training and the test, and what topics were left out. how our proposed model compares to other "state-of-the-art" methods in terms of accuracy and F1-score. We now all know that we know how to tell if news is real or not. On the other hand, I want to bring the following to your attention:

It is very hard, if not impossible, to figure out what an article is about just by looking at its headline. It's nothing like what you think!! After thinking about how the news story fits into the bigger picture, we came up with a much better idea. In a very short amount of time, the success rate went from 94% to over 99.0%. It's a huge step in the right direction. Third, by taking into account both the name and the location, we have reached the most precise level that is even remotely possible. It is my plan to work on the implementation of other models that will lead to even better results.

Please let me make the following points about the information:

Since the data that was collected was already pretty clean, it only needed a little bit of cleaning. A few things have been changed so that they better meet my needs.

Even though we will try to make this better in the future, we know that we need a lot more data right now to train the model.

Indicators and Measures of Performance We use the evaluation metrics of accuracy (A), precision (P), recall (R), and F1-score to compare and analyse our model (F). In order to do the math for Precision and Recall, we need equations 1 and 2. On the other hand, the F1-score is a mathematical method that tries to find a balance between recall and accuracy.

$$P = \frac{TP}{TP + FP}$$

The accuracy of a classifier is measured by how many positive class values it correctly labels out of the total number of those values (including true and false). It shows how realistic the model is.

$$R = \frac{TP}{TP + FN}$$

The recall rate is found by dividing the number of times a positive class was correctly identified by the sum of the number of times the positive class was correctly identified and the number of times the negative class was wrongly identified. It gives information about how complete the model is.

$$F1 = 2 * \frac{precision.recall}{precision + recall}$$

For each category, there is something called an F1 score that is used to measure how accurate a model is. The F1-score measure is often used when there is an imbalance in the data set. We use F1-score as an evaluation metric to show that the proposed model is complete in terms of class-wise accuracy because the dataset used by FNC-1 is just as wildly unbalanced as the dataset used by FNC-1.

Table-1 Comparative Analysis

Model	Precision	Recall	F1	Accuracy
LR	78.543	72.431	72.432	80.323
RF	80.235	75.345	73.456	82.344
PCA	81.345	76.675	74.345	85.654
SVM	85.543	84.345	81.234	88.345
SVM Kernel	90.453	88.345	78.345	91.234
Proposed Model	94.345	90.345	80.345	94.345

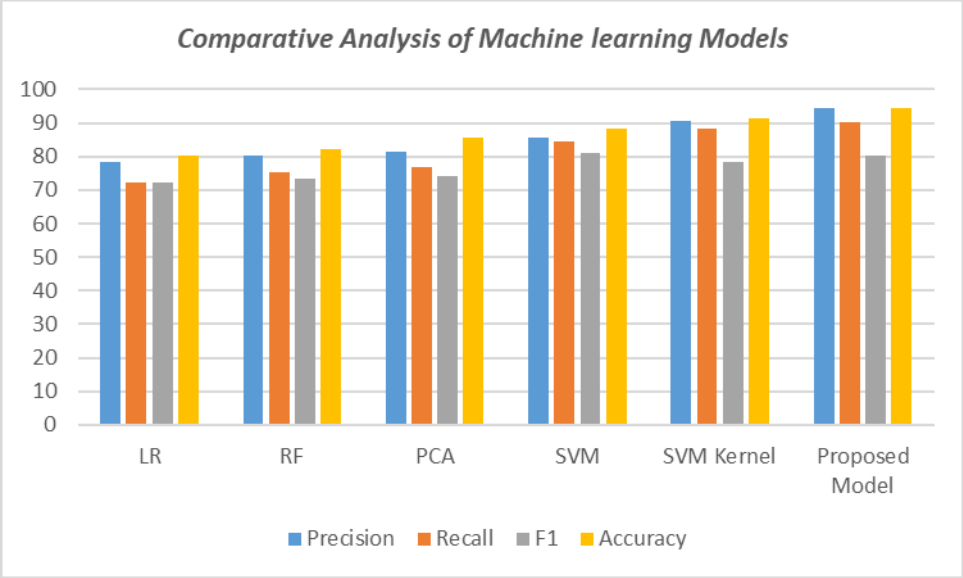

Figure 2-Comparative Analysis of Machine learning Models

Conclusion

Because it's so easy to get on the internet now, false information is getting around faster than ever. A lot of people can always get on the Internet and use social media in different ways. When you use these ways to share news, you are not limited in what you can say. But a loud but small group of people have started using these platforms to spread false information about other people or institutions. This is one of the worst things that can happen to someone or a business's reputation. If false information about a political party gets out to the public, it could change how people feel about that party. It is important to do research and come up with ways to spot fake news. Classifiers that are based on machine learning can be used to help find hoaxes, in addition to their many other

uses. Before they can be used in production, classifiers need to be taught the right way to look at a certain set of data, which is called the training data set. After that, these classifiers will be able to figure out on their own when a story isn't true. This study of the research literature focuses on supervised machine learning classifiers, which are defined as those that must be trained using data that has been labelled. There isn't a lot of easily accessible labelled data that can be used to train classifiers to spot fake news. In the future, it's possible that research will be done on unsupervised machine learning classifiers to help find fake news.

Reference

[1]. Gupta et al., "Combating Fake News: Stakeholder Interventions and Potential Solutions," in IEEE Access, vol. 10, pp. 78268-78289, 2022, doi: 10.1109/ACCESS.2022.3193670.

[2]. D. Rohera et al., "A Taxonomy of Fake News Classification Techniques: Survey and Implementation Aspects," in IEEE Access, vol. 10, pp. 30367-30394, 2022, doi: 10.1109/ACCESS.2022.3159651.

[3]. E. Elsaeed, O. Ouda, M. M. Elmogy, A. Atwan and E. El-Daydamony, "Detecting Fake News in Social Media Using Voting Classifier," in IEEE Access, vol. 9, pp. 161909-161925, 2021, doi: 10.1109/ACCESS.2021.3132022.

[4]. H. Saleh, A. Alharbi and S. H. Alsamhi, "OPCNN-FAKE: Optimized Convolutional Neural Network for Fake News Detection," in IEEE Access, vol. 9, pp. 129471-129489, 2021, doi: 10.1109/ACCESS.2021.3112806.

[5]. M. Narra et al., "Selective Feature Sets Based Fake News Detection for COVID-19 to Manage Infodemic," in IEEE Access, vol. 10, pp. 98724-98736, 2022, doi: 10.1109/ACCESS.2022.3206963.

[6]. M. Umer, Z. Imtiaz, S. Ullah, A. Mehmood, G. S. Choi and B. -W. On, "Fake News Stance Detection Using Deep Learning Architecture (CNN-LSTM)," in IEEE Access, vol. 8, pp. 156695-156706, 2020, doi: 10.1109/ACCESS.2020.3019735.

[7]. P. K. Verma, P. Agrawal, I. Amorim and R. Prodan, "WELFake: Word Embedding Over Linguistic Features for Fake News Detection," in IEEE Transactions on Computational Social Systems, vol. 8, no. 4, pp. 881-893, Aug. 2021, doi: 10.1109/TCSS.2021.3068519.

[8]. P. Wei, F. Wu, Y. Sun, H. Zhou and X. -Y. Jing, "Modality and Event Adversarial Networks for Multi-Modal Fake News Detection," in IEEE Signal Processing Letters, vol. 29, pp. 1382-1386, 2022, doi: 10.1109/LSP.2022.3181893.

[9]. K. A. Qureshi, R. A. S. Malick, M. Sabih and H. Cherifi, "Complex Network and Source Inspired COVID-19 Fake News Classification on Twitter," in IEEE Access, vol. 9, pp. 139636-139656, 2021, doi: 10.1109/ACCESS.2021.3119404.

[10]. N. O. Bahurmuz, G. A. Amoudi, F. A. Baothman, A. T. Jamal, H. S. Alghamdi and A. M. Alhothali, "Arabic Rumor Detection Using Contextual Deep Bidirectional Language Modeling," in IEEE Access, vol. 10, pp. 114907-114918, 2022, doi: 10.1109/ACCESS.2022.3217522.

[11]. D. S. Abdelminaam, F. H. Ismail, M. Taha, A. Taha, E. H. Houssein and A. Nabil, "CoAID-DEEP: An Optimized Intelligent Framework for Automated Detecting COVID-19 Misleading Information on Twitter," in IEEE Access, vol. 9, pp. 27840-27867, 2021, doi: 10.1109/ACCESS.2021.3058066.

[12]. G. Sansonetti, F. Gasparetti, G. D'aniello and A. Micarelli, "Unreliable Users Detection in Social Media: Deep Learning Techniques for Automatic Detection," in IEEE Access, vol. 8, pp. 213154-213167, 2020, doi: 10.1109/ACCESS.2020.3040604.

[13]. M. K. Elhadad, K. F. Li and F. Gebali, "Detecting Misleading Information on COVID-19," in IEEE Access, vol. 8, pp. 165201-165215, 2020, doi: 10.1109/ACCESS.2020.3022867.

[14]. Palani, B., Elango, S. & Viswanathan K, V. CB-Fake: A multimodal deep learning framework for automatic fake news detection using capsule neural network and BERT. Multimed Tools Appl 81, 5587–5620 (2022). https://doi.org/10.1007/s11042-021-11782-3.

[15]. Kaliyar, R.K., Goswami, A. & Narang, P. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimed Tools Appl* **80**, 11765–11788 (2021). https://doi.org/10.1007/s11042-020-10183-2.

[16]. Kausar, N., AliKhan, A. & Sattar, M. Towards better representation learning using hybrid deep learning model for fake news detection. Soc. Netw. Anal. Min. 12, 165 (2022). https://doi.org/10.1007/s13278-022-00986-6.

[17]. Zervopoulos, A., Alvanou, A.G., Bezas, K. et al. Deep learning for fake news detection on Twitter regarding the 2019 Hong Kong protests. Neural Comput & Applic 34, 969–982 (2022). https://doi.org/10.1007/s00521-021-06230-0.

[18]. Agarwal, A., Mittal, M., Pathak, A. et al. Fake News Detection Using a Blend of Neural Networks: An Application of Deep Learning. SN COMPUT. SCI. 1, 143 (2020). https://doi.org/10.1007/s42979-020-00165-4.

[19]. Kaliyar, R.K., Goswami, A. & Narang, P. DeepFakE: improving fake news detection using tensor decomposition-based deep neural network. J Supercomput 77, 1015–1037 (2021). https://doi.org/10.1007/s11227-020-03294-y.

[20]. Fayaz, M., Khan, A., Bilal, M. et al. Machine learning for fake news classification with optimal feature selection. Soft Comput 26, 7763–7771 (2022). https://doi.org/10.1007/s00500-022-06773-x.

[21]. Hanshal, O.A., Ucan, O.N. & Sanjalawe, Y.K. Hybrid deep learning model for automatic fake news detection. *Appl Nanosci* (2022). https://doi.org/10.1007/s13204-021-02330-4.

[22]. Li, X., Lu, P., Hu, L. et al. A novel self-learning semi-supervised deep learning network to detect fake news on social media. Multimed Tools Appl 81, 19341–19349 (2022). https://doi.org/10.1007/s11042-021-11065-x.

[23]. Kaliyar, R.K., Goswami, A. & Narang, P. EchoFakeD: improving fake news detection in social media with an efficient deep neural network. Neural Comput & Applic 33, 8597–8613 (2021). https://doi.org/10.1007/s00521-020-05611-1

[24]. K, S., Thilagam, P.S. Multi-layer perceptron based fake news classification using knowledge base triples. Appl Intell (2022). https://doi.org/10.1007/s10489-022-03627-9.

[25]. Javed, M.S., Majeed, H., Mujtaba, H. et al. Fake reviews classification using deep learning ensemble of shallow convolutions. J Comput Soc Sc 4, 883–902 (2021). https://doi.org/10.1007/s42001-021-00114-y.

[26]. Meesad, P. Thai Fake News Detection Based on Information Retrieval, Natural Language Processing and Machine Learning. SN COMPUT. SCI. 2, 425 (2021). https://doi.org/10.1007/s42979-021-00775-6

[27]. Cartwright, B., Frank, R., Weir, G. et al. Detecting and responding to hostile disinformation activities on social media using machine learning and deep neural networks. Neural Comput & Applic 34, 15141–15163 (2022). https://doi.org/10.1007/s00521-022-07296-0

[28]. Iftikhar Ahmad et al. “Fake news detection using machine learningensemble methods”. In: Complexity 2020 (2020).

[29]. sMonther Aldwairi and Ali Alwahedi. “Detecting fake news insocial media networks”. In: Procedia Computer Science 141 (2018),pp. 215–222