International Journal of Current Research and Techniques (IJCRT)
E-ISSN: 2349-3194 (Online) | P-ISSN: 2348-4446 (Print)
UGC Approved Journal No: 47722 [No: 2616]
DOI: https://doi.org/10.61359/IJCRT2024050022

Enhanced Fake News Detection with the aid of Improved Spider Monkey Optimization-based optimal Feature Selection and Deep Neural Network

Salim G. Shaikh, Associate Professor, Department of Computer Engineering, Kalsekar Technical Campus, Panvel India
Nishant N. Pachpor, Assistant Professor, IIMS, Chinhcwad Pune, India
Mukhtar Ansari, Assistant Professor, Department of Computer Engineering, Kalsekar Technical Campus, Panvel India

Manuscript Received: Oct 15, 2024; Published: Oct 28, 2024

Abstract

Fake news has become a significant problem in recent years, leading to widespread misinformation and public manipulation. This research focuses on developing an effective fake news detection model using advanced machine learning and deep learning techniques. Existing methodologies face challenges such as poor performance with large datasets, noise, and limited generalization. The proposed solution integrates pre-processing, feature extraction, optimal feature selection via Spider Monkey Optimization (SMO), and a deep neural network (OAF-DNN) with optimized activation functions. The model's performance will be validated using publicly available datasets and analyzed through various evaluation metrics. This study aims to enhance the accuracy, precision, and detection of fake news.

Keywords: Spider Monkey Optimization (SMO), Deep Neural Network (DNN), Fake news detection(FND), Social media, Text classification.

Introduction

Over the past several years, fake news dissemination has become a major problem. Fake news is defined by the New York Times as “made-up stories are written to deceive”, and they are published in formats similar to those used by traditional news agencies [9]. Fake news has been identified for contributing to increased political polarization and partisan conflict in recent times. Recent examples included the controversy created during the 2016 presidential campaign for the United States [9] and Indian Airstrike in Balakot in 2019. Fake news is a text classification issue with a straight forward proposition [10]. The potential for dissemination, acceptance, and destruction of fake news poses them as one of the greatest threats to the concept of logical truth. Since the popularization of the spread of fake news, there has been a growing joint effort by the academic community to research and develop approaches capable of analyzing, detecting and intervening in the performance of misleading content. Scientific evidence has already revealed the human vulnerability in distinguishing real from false facts, while the capacity for human differentiation reduces up to a random probability of approximately 54% correctness. Furthermore, the fight against fake news renders the social network and data consumption problems inseparable. By spreading malicious content, a user is wasting network and processing resources and undermining the credibility of the service provided. In turn, fake news hamper the Quality of Trust (QoT) applied to news distribution, that is, how much a user trusts in the content of a particular source [11] [12] [13] [14] [15] [16].

Over the past decade, we have witnessed the development of fake news detection technologies, mainly grouping into cue and feature-based methods [17] [18] [19], which can be employed to distinguish fake news contents from true news contents by designing a set of linguistic cues that are informative of the content veracity, and linguistic analysis based methods [20] [21], which can be applied to distinguish fake from true news by exploiting differences in writing style, language, and sentiment. Such methods do not require task-specific, hand engineered cue sets and rely on automatically extracting linguistic features from the text. Unfortunately, variations in linguistic cues implies that a new cue set must be designed for a new situation, making it hard to generalize cue and feature engineering methods across topics and domains. Linguistic analysis methods, although better than cue-based methods, still do not fully extract and exploit the rich semantic and syntactic information in the content. Neural network is an attracted machine learning model that can learn the nonlinear mapping from data, especially for deep model. They have been employed for automatic detection of fake news [22] [23] [24] and show an impressed practical performance. However, even with sophisticated feature extraction of deep learning methods, fake news detection remains to be a challenge, primarily because the content is crafted to resemble the truth in order to deceive readers, and without fact-checking or additional information, it is often hard to determine veracity by text analysis alone [25].

Related Works

In 2020, Oliveira et al. [1] have proposed a computational-stylistic analysis based on natural language processing, efficiently applying machine learning algorithms to detect fake news in texts extracted from social media. The analysis considers news from Twitter, from which approximately 33,000 tweets were collected, assorted between real and proven false. In assessing the quality of detection, 86% accuracy, and 94% precision stand out even employing a dimensional reduction to one-sixth of the number of original features. The developed approach introduced a minimum overhead, while it has the potential of providing a high confidence index on discriminating fake from real news.

In 2020, Huang and Chen [2] have proposed a fake news detection system using a deep learning model. First, news articles were preprocessed and analyzed based on different training models. Then, an ensemble learning model combining four different models called embedding LSTM, depth LSTM, LIWC CNN, and N-gram CNN was proposed for fake news detection. Besides, to achieve higher accuracy in fake news detection, the optimized weights of the ensemble learning model were determined using the Self-Adaptive Harmony Search (SAHS) algorithm. In the experiments, the authors have verified that the proposed model was superior to the state-of-the-art methods, with the highest accuracy of 99.4%. Furthermore, the investigation of the cross-domain intractability issue was done and achieved the highest accuracy of 72.3%. Finally, the authors believed there was still room for improving the ensemble learning model in addressing the cross-domain intractability issue.

In 2020, Kaliyar et al. [3] have proposed a deep convolutional neural network (FNDNet) for fake news detection. Instead of relying on hand-crafted features, the proposed model (FNDNet) was designed to automatically learn the discriminatory features for fake news classification through multiple hidden layers built in the deep neural network. The authors created a deep Convolutional Neural Network (CNN) to extract several features at each layer. The authors compared the performance of the proposed approach with several baseline models. Benchmarked datasets were used to train and test the model, and the proposed model achieved state-of-the-art results with an accuracy of 98.36% on the test data. Various performance evaluation parameters such as Wilcoxon, false positive, true negative, precision, recall, F1, and accuracy, etc. were used to validate the results. These results demonstrated significant improvements in the area of fake news detection as compared to existing state-of-the-art results and affirm the potential of our approach for classifying fake news on social media. This research will assist researchers in broadening the understanding of the applicability of CNN-based deep models for fake news detection.

In 2020, Xu et al. [4] have characterized hundreds of popular fake and real news measured by shares, reactions, and comments on Facebook from two perspectives: domain reputations and content understanding. The presented domain reputation analysis revealed that the Web sites of the fake and real news publishers exhibit diverse registration behaviors, registration timing, domain rankings, and domain popularity. In addition, fake news tends to disappear from the Web after a certain amount of time. The content characterizations on the fake and real news corpus suggest that simply applying term frequency-inverse document frequency (tf-idf) and Latent Dirichlet Allocation (LDA) topic modeling was inefficient in detecting fake news, while exploring document similarity with the term and word vectors was a very promising direction for predicting fake and real news. To the best of our knowledge, this was the first effort to systematically study domain reputations and content characteristics of fake and real news, which has provided key insights for effectively detecting fake news on social media.

In 2020, Ozbay and Alatas [5] have proposed a two-step method for identifying fake news on social media, focused on fake news. In the first step of the method, a number of pre-processing was applied to the data set to convert un-structured data sets into the structured data set. The texts in the data set containing the news were represented by vectors using the obtained TF weighting method and Document-Term Matrix. In the second step, twenty-three supervised artificial intelligence algorithms have been implemented in the data set transformed into the structured format with the text mining methods. An experimental evaluation of the twenty-three intelligent classification methods has been performed within existing public data sets and these classification models have been compared depending on four evaluation metrics.

In 2020, Kaliyar et al. [6] have considered the content of the news article and the existence of echo chambers in the social network for fake news detection. A tensor representing social context was formed by combining the news, user and community information. The news content was fused with the tensor, and coupled matrix-tensor factorization was employed to get a representation of both news content and social context. The proposed method has been tested on a real-world dataset: BuzzFeed. The factors obtained after decomposition have been used as features for news classification. An ensemble machine learning classifier (XGBoost) and a deep neural network model (DeepFakE) were employed for the task of classification. The proposed model (DeepFakE) outperformed with the existing fake news detection methods by applying deep learning on combined news content and social context-based features as an echo-chamber.

In 2020, Li et al. [7] have proposed multi-level convolutional neural network (MCNN), which introduced the local convolutional features as well as the global semantics features, to effectively capture semantic information from article texts which can be used to classify the news as fake or not. The authors later employed a method of calculating the weight of sensitive words (TFW), which has shown their stronger importance with their fake or true labels. Finally, MCNN-TFW was developed, a multiple-level convolutional neural network-based fake news detection system, which was combined to perform fake news detection in that MCNN extracts article representation and WS calculated the weight of sensitive words for each news. Extensive experiments have been done on fake news detection in cultural communication to compare MCNN-TFW with several state-of-the-art models, and the experimental results have demonstrated the effectiveness of the proposed model.

In 2020, Reddy et al. [8] have discussed approaches to detection of fake news using only the features of the text of the news, without using any other related metadata. The authors have observed that a combination of stylometric features and text-based word vector representations through ensemble methods predicted fake news with an accuracy of up to 95.49%.

Problem Definition

Fake news denotes a category of daily mail that actively spreads lies or falsifications that propagate across both conventional print news outlets and online social media. As the "Great moon hoax" published in the year 1835, fake news will be present for longer time. In the online environment, fake news for discrete political and commercial reasons is found to be more due to the growing technology of online social networks in the earlier years. There are some advantages and disadvantages with the existing fake news detection models as shown in Table 1. Among them, one class SVM [1] is used to learn single class for classifier training, and it is employed for deriving the decision hyperplane for anomaly detection. However, it is not appropriate for huge datasets, and it doesn't perform well for noisy data. SAHS [2] determines of optimized weights of ensemble learning model, and it has high accuracy in detecting fake news. But, it won't detect the fake news in an early stage, and the main text decreases the chance in discriminating the fake news from real news. Deep CNN [3] is employed for extracting many features in each layer, and it is less vulnerable to overfitting problem. Yet, there are some conflicts like it doesn't encode the object's position and orientation, and it requires being spatially invariant to the input data. TF-IDF [4] is employed for analyzing the similarity and dissimilarity of real and fake news on the most significant terms of news articles, and it has high efficiency. Still, it calculates the document similarity directly in the word count space that slows down the large vocabularies, and it doesn't have the ability to capture semantics. Supervised artificial intelligence algorithm [5] is used to apply on the fake news dataset, and is employed for estimating the output with minimum error rate from the input data. However, there are some disadvantages such as ensemble approaches and various feature extraction models need to be integrated for best performance, and it doesn't have the ability to handle some of the complicated tasks. Deep Neural Network [6] is utilized for modelling the combined representation of fake news detection, and it has attained the best classification results. But, it needs more amount of data for attaining best performance, and it is quite expensive because of complex data models. MCNN [7] is employed for developing the local Convolutional features and global semantic features for efficient semantic data capturing, and it is utilized for extracting the article representation. But, it needs to be implemented in wide range of applications, and is very slow because of maxpooling operation. Ensemble Methods [8] has the ability to predict fake news with more accuracy, and in order to attain voting results, ensemble models are applied on the accumulation of skip gram, writeprints, and stylometric features. Yet, it doesn't have the ability to interpret the model, and it requires more computation and design time. Hence, it is concluded that the above mentioned challenges are helpful for developing a new model for effective detection of fake news.

International Journal of Current Research and Techniques (IJCRT)
E-ISSN: 2349-3194 (Online) | P-ISSN: 2348-4446 (Print)
UGC Approved Journal No: 47722 [No: 2616]
DOI: https://doi.org/10.61359/IJCRT2024050022

Table 1: Features and challenges of existing fake news detection models

Author [citation]	Methodology	Features	Challenges
Oliveira *et al*. [1]	One class SVM	• It is used to learn single class for classifier training. • It is employed for deriving the decision hyperplane for anomaly detection.	• It is not appropriate for huge datasets. • It doesn't perform well for noisy data.
Huang and Chen [2]	SAHS	• The determination of optimized weights of ensemble learning model. • It has high accuracy in detecting fake news.	• It won't detect the fake news in an early stage. • The main text decreases the chance in discriminating the fake news from real news.
Kaliyar *et al*. [3]	Deep CNN	• It is employed for extracting many features in each layer. • It is less vulnerable to overfitting problem.	• It doesn't encode the object's position and orientation. • It requires being spatially invariant to the input data.
Xu *et al*. [4]	TF-IDF	• It is employed for analyzing the similarity and dissimilarity of real and fake news on the most significant terms of news articles. • It has high efficiency.	• It calculates the document similarity directly in the word count space that slows down the large vocabularies. • It doesn't have the ability to capture semantics.
Ozbay and Alatas [5]	supervised artificial intelligence algorithm	• It is used to apply on the fake news dataset. • It is employed for estimating the output with minimum error rate from the input data.	• Ensemble approaches and various feature extraction models need to be integrated for best performance. • It doesn't have the ability to handle some of the complicated tasks.
Kaliyar *et al*. [6]	Deep Neural Network	• It is utilized for modelling the combined representation of fake news detection. • It has attained the best classification results.	• It needs more amount of data for attaining best performance. • It is quite expensive because of complex data models.
Li *et al*. [7]	MCNN	• It is employed for developing the local Convolutional features and global semantic features for efficient semantic data capturing. • It is utilized for extracting the article representation.	• Needs to be implemented in wide range of applications. • It is very slow because of maxpooling operation.
Reddy *et al*. [8]	Ensemble Methods	• It has the ability to predict fake news with more accuracy. • In order to attain voting results, ensemble models are applied on the accumulation of skip gram, writeprints, and stylometric features.	• It doesn't have the ability to interpret the model. • It requires more computation and design time.

Research Objectives

The objectives of this research work are based on the following points.

- To undergo a critical review on face news detection and states the features and challenges of the state-of-the art methodologies.

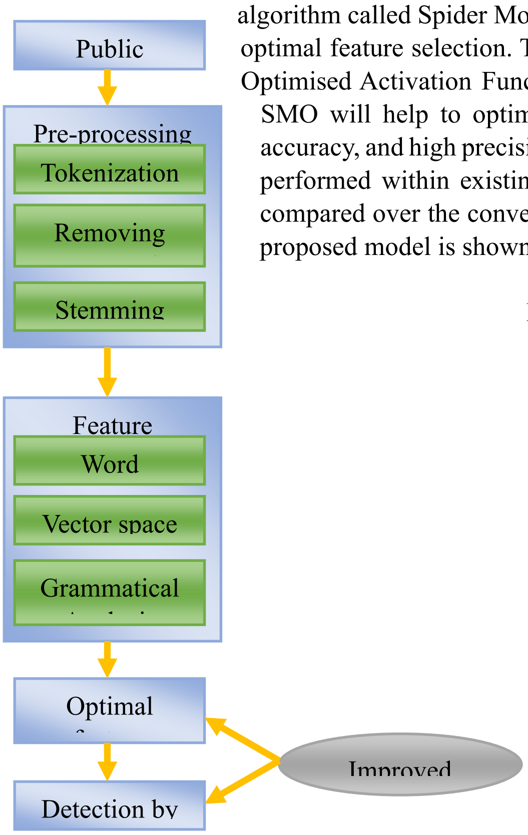
International Journal of Current Research and Techniques (IJCRT)
E-ISSN: 2349-3194 (Online) | P-ISSN: 2348-4446 (Print)
UGC Approved Journal No: 47722 [No: 2616]
DOI: https://doi.org/10.61359/IJCRT2024050022

- To introduce the new feature extraction methods that could positively helps to determine difference between the real and fake news.
- To develop a novel feature selection approach that could enables the machine learning algorithm to train faster and reduces the complexity.
- To implement the new optimized deep learning model for assisting the categorization of real and fake news effectively with diverse architectural improvements.
- To improvise the existing meta-heuristic algorithms for enhancing the performance of fake news detection, thus influencing the selection of best solution to attain the best detection outcome.
- To validate the performance of the proposed and conventional models through valuable analysis in terms of diverse performance measures.

Research Methodology and Proposed Model

Automatic text summarization offers an efficient solution to access the ever-growing amounts of both Fake news has recently leveraged the power and scale of online social media to effectively spread misinformation which not only erodes the trust of people on traditional presses and journalisms, but also manipulates the opinions and sentiments of the public. Detecting fake news is a daunting challenge due to subtle difference between real and fake news. With deceptive words, online social network users can get infected by this online fake news easily, which has brought about tremendous effects on the offline society already. An important goal in improving the trustworthiness of information in online social networks is to identify the fake news timely. However, fake news detection remains to be a challenge, primarily because the content is crafted to resemble the truth in order to deceive readers, and without fact-checking or additional information, it is often hard to determine veracity by text analysis alone. The main intent of this proposal is to plan for the effective and reliable fake news detection model using the intelligent technology. The proposed model covers diverse phases like (a) Pre-processing, (b) Feature extraction, (c) optimal feature selection, and (d) detection. The publically available benchmark dataset will be gathered from different sources, and applied with pre-processing initially. The methods such as tokenization, removing stop words, and stemming will be used for pre-processing of data. Further, the feature extraction will be done that involves word embedding using word2vect, Vector Space Model (VSM) using Term Frequency-Inverse Document Frequency (TF-IDF), and grammatical analysis using mean, Q25, Q50, Q75, Max, Min, and standard deviation. Since the length of extracted features are long, feature extraction is induced, which enables the machine learning algorithm to train faster and reduces the complexity. Here, the improved meta-heuristic algorithm called Spider Monkey Optimization (SMO) [26] will be used for performing the optimal feature selection. The detection of real or fake news will be accomplished by the Optimised Activation Function-based Deep Neural Network (OAF-DNN), in which the SMO will help to optimize the activation function in order to attain high detection accuracy, and high precision. An experimental evaluation of the proposed model has been performed within existing public data sets and these classification models have been compared over the conventional methods depending on diverse evaluation metrics. The proposed model is shown in Fig. 1.

Figure 1: Proposed fake news detection model



Expected Results

The proposed fake news detection model will be implemented in python, and the experimental analysis will be carried out. Here, the performance of the proposed model will be compared over the conventional models by analysing the Type I and Type II performance measures. Here, Type I measures are positive measures like Accuracy, Sensitivity, Specificity, Precision, Negative Predictive Value (NPV), F1Score and Mathews correlation coefficient (MCC), and Type II measures are negative measures like False positive rate (FPR), False negative rate (FNR), and False Discovery Rate (FDR).

Conclusion

The rise of fake news, particularly through social media, has created a significant challenge for both individuals and institutions. Despite the advancements in fake news detection models, many approaches still struggle with generalization, efficiency, and accuracy. This research introduces a novel deep learning-based fake news detection model incorporating Spider Monkey Optimization (SMO) for optimal feature selection and performance enhancement. By utilizing innovative feature extraction methods and improving existing algorithms, this model offers a promising solution for detecting fake news more effectively. The experimental analysis demonstrates the proposed model's superiority in handling complex data and provides higher detection accuracy compared to conventional methods.

References

[1] 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 Processing Letters, vol. 27, pp. 1250-1254, 2020.

[2] Yin-Fu Huang, and Po-Hong Chen, "Fake news detection using an ensemble learning model based on Self-Adaptive Harmony Search algorithms", Expert Systems with Applications, vol. 159, 30 November 2020.

[3] Rohit Kumar Kaliyar, Anurag Goswami, Pratik Narang, and Soumendu Sinha, "FNDNet – A deep convolutional neural network for fake news detection", Cognitive Systems Research, vol. 61, pp. 32-44, June 2020.

[4] K. Xu, F. Wang, H. Wang and B. Yang, "Detecting fake news over online social media via domain reputations and content understanding," Tsinghua Science and Technology, vol. 25, no. 1, pp. 20-27, Feb. 2020.

[5] Feyza Altunbey Ozbay, and Bilal Alatas, "Fake news detection within online social media using supervised artificial intelligence algorithms", Physica A: Statistical Mechanics and its Applications, vol. 540, 15 February 2020.

[6] Rohit Kumar Kaliyar, Anurag Goswami & Pratik Narang, "DeepFakE: improving fake news detection using tensor decomposition-based deep neural network", The Journal of Supercomputing, 2020.

[7] Qian Li, Qingyuan Hu, Youshui Lu, Yue Yang & Jingxian Cheng, "Multi-level word features based on CNN for fake news detection in cultural communication", Personal and Ubiquitous Computing, vol. 24, pp. 259–272, 2020.

[8] Harita Reddy, Namratha Raj, Manali Gala & Annappa Basava, "Text-mining-based Fake News Detection Using Ensemble Methods", International Journal of Automation and Computing, vol. 17, pp. 210-221, 2020.

[9] Jeff Z. Pan, Siyana Pavloval, Chenxi Li1, Ningxi Li1, Yangmei Li1, and Jinshuo Liu, "Content Based Fake News Detection Using Knowledge Graphs", International Semantic Web Conference, pp. 669-683, 2018.

[10] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H, "Fake news detection on social media: A data mining perspective", ACM SIGKDD Explorations Newsletter, September 2017.

[11] Soroush Vosoughi, Deb Roy, and Sinan Aral, "The spread of true and false news online", Science, vol. 359, no. 6380, pp. 1146–1151, 2018.

[12] X. Zhou and R. Zafarani, "Fake news: A survey of research, detection methods, and opportunities," arXiv preprint arXiv:1812.00315, 2018.

[13] W. Y. Wang, ""Liar, liar pants on fire": A new benchmark dataset for fake news detection," in Annual Meeting of the Association for Computational Linguistics - ACL 2017, 2017.

International Journal of Current Research and Techniques (IJCRT)
E-ISSN: 2349-3194 (Online) | P-ISSN: 2348-4446 (Print)
UGC Approved Journal No: 47722 [No: 2616]
DOI: https://doi.org/10.61359/IJCRT2024050022

[14] V. L. Rubin, “On deception and deception detection: Content analysis of computer-mediated stated beliefs,” in Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem, vol. 47, pp. 32, 2010.
[15] V. Rubin, N. Conroy, Y. Chen, and S. Cornwell, “Fake news or truth? using satirical cues to detect potentially misleading news,” in Proceedings of the Second Workshop on Computational Approaches to Deception Detection, pp. 7–17, 2016.
[16] G. Liu, Y. Wang, and M. A. Orgun, “Quality of trust for social trust path selection in complex social networks,” in Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems, vol. 1, pp. 1575–1576, 2010.
[17] Bogaard G, Meijer EH, Vrij A, Merckelbach H, "Scientific Content Analysis (SCAN) cannot distinguish between truthful and fabricated accounts of a negative event", vol. 7, pp. 243, 2016.
[18] Galit Nahari, Aldert Vrij & Ronald P. Fisher, "Does the Truth Come Out in the Writing? SCAN as a Lie Detection Tool", Law and Human Behavior, 2011.
[19] Castillo C, Mendoza M, Poblete B, "Information credibility on Twitter", Proceedings of the 20th International Conference on World Wide Web, 2011.
[20] Ott M, Cardie C, Hancock JT, "Negative deceptive opinion spam", In HLT-NAACL, pp. 497-501, 2013.
[21] Myle Ott, Yejin Choi, Claire Cardie, Jeffrey T. Hancock, "Finding Deceptive Opinion Spam by Any Stretch of the Imagination", arXiv.org > cs > arXiv:1107.4557, 22 Jul 2011.
[22] Blunsom P, Grefenstette P, Kalchbrenner N, "A Convolutional Neural Network for Modelling Sentences", Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 655–665, June 2014.
[23] Kimura M, Saito K, Motoda H, "Efficient estimation of influence functions for SIS model on social networks", IJCAI, pp. 2046-2051, 2009.
[24] Yih W-t, He X, Meek C, "Semantic parsing for single-relation question answering", In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, vol. 2, pp. 643-648, 2014.
[25] Yoon Kim, "Convolutional Neural Networks for Sentence Classification", Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1746-1751, 2014.
[26] Vani Agrawal, and Ratika Rastogi, and D. C. Tiwari, "Spider Monkey Optimization: a survey", International Journal of System Assurance Engineering and Management, vol. 9, no. 4, pp. 929–941, August 2018.