iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** V **Month of publication:** May 2024

DOI: https://doi.org/10.22214/ijraset.2024.61528

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

Data & Semantic Analysis Fake News Detection Based Domain-Specific Auto Encoder Learning

Dr.Karthik Elangovan[1], Bipul Kumar Singh[2], Varun R[3], Rehan Ahmed T[4]
[1]*Assistant Professor Department of Computer Science and Engineering SRM Institute of Science and Technology Ramapuram Chennai*
[2, 3, 4]*B.Tech CSE with Specialization in Cyber Security, SRM University, India*

Abstract: ***Efforts to create technology for automatically detecting fake news are actively underway as the spread of misinformation on social media continues to escalate. However, most of these approaches primarily concentrate on the linguistic and structural aspects of fake news, such as identifying sources or authors, message length, and the frequency of negative language. In contrast, our study introduces a phony news detection model leveraging machine learning that incorporates user behaviors, news, and social capital-based social network dynamics. We used the XG Boost model to evaluate each feature's significance and identify the critical elements affecting the identification of fake news to capture the variables related to the spread of phony news entirely. We implemented SVM, RF, LR, CART, and NNET, well-known machine learning classification models, using these identified factors and assessed how well they detected bogus news. This work used a cross-validation process to prevent overfitting and to generalize the established models. Additionally, the predicted accuracy of the established models was compared. The RF model had the best forecast accuracy, about 94%, while the NNET had the worst performance rate, about 92.1%. As disinformation is created and disseminated with growing sophistication, the results of this study should improve the effectiveness of false news detection systems.***

I. INTRODUCTION

Initially developed to enhance user satisfaction, this personalised content-recommended system has recently evolved into a significant conduit for disseminating fake news. [1]. These are made to constantly display material in the feed that matches what the user has seen or engaged with through "Likes." In our model, comments receive the same treatment regardless of the veracity of the story. [2].. Put differently, the system must repeatedly suggest information similar to that of users exposed to bogus news. They even consciously alter how pointless stuff appears on social media. [3]. Previous research has concentrated on identifying or detecting bogus news based on linguistic or compositional traits. They have distinguished fake news based on factors like the article's length and the presence of an identifiable author and source [4, 5, 6, 7]. This approach assumes that real and fake news differ in specific language or structural aspects. Reflecting the traits of users who agree to disseminate false information, as well as the characteristics of the social media platforms on which it proliferates, is challenging

Since Chat GPT (Generative Pre-trained Transformer) can identify non-uncomfortable style elements, like those created by minimal AI, it is now impossible to guarantee that fake news can be correctly recognized in the past. Nowadays, users confuse news from Chat GPT for news written by a professional journalist in a matter of seconds due to how easy it is to create false news that seems authentic. [48]. As a result, it's essential to recognise fraudulent information in a different way than before. In this study, we finally provide a detection model that takes into account the visual characteristics of the material, the characteristics of the persons who produce and spread false news, and the networks that spread it.. Furthermore, fake news may be produced with greater sophistication thanks to modern technology that uses artificial intelligence (AI) to quickly and efficiently create content that looks precisely like actual news. AI-powered Twitter bots (such as the EAI Twitter Bot) can create hundreds of user accounts that can support or oppose any information that appears to be authentic news, regardless of whether it is phony. The bot can take control of the target. As an outcome, it is getting progressively more challenging to identify carefully manufactured fake news based just on its surface characteristics. To counter the current false news detection method that relies on linguistic features, a more diverse approach to social media user profiles and networks is required. The goal of this work is to improve the detection accuracy of fake news by addressing a restriction that was disregarded in earlier research: the characteristics of information recipients. As a result, we create a model for detecting false news by taking into account different content characteristics, individuals, and the social media and networks where fake news is created and shared. Identifying the appropriate characteristics is crucial when recognising fake news, as there are various explanatory variables to consider.

The XGBoost (Extreme Gradient Boosting) method is used first to determine the explanatory variable. By exploiting well-selected explanatory variables, we want to improve the accuracy of identifying fake news by creating an ideal detection model with XGBoost.Classification and Regression Trees (CART), Logistic Regression (LR), Support Vector Machine (SVM), Neural Network (NNET), and Random Forest (RF) are all used in the field of classification.) are the five machine learning approaches that are used in its construction. After comparing their performance rates, the algorithm that can anticipate bogus news the best is the winner.

II. RELATED WORK

Current study demonstrates the range of techniques employed by scientists to develop machine learning models for the detection of false news and the possibility that these models may improve the accuracy of news verification. The development of algorithms for detecting false news is difficult, and like any other technology, they occasionally make mistakes in identifying fake news.. Fake news can quickly spread throughout social media platforms and significantly impact public opinion and behaviour if the system fails to identify it. For instance, false information that had been taken as gospel in the first place caused hundreds of individuals to pass away in Iran after consuming methanol to treat COVID-19 [47]. Moreover, inaccuracies in fake news detection algorithms may result in erroneous allegations or incorrect group or person identifications. A valid news piece may be wrongly labeled as fake news by a system, which may result in charges of bias or censorship directed towards the news organization that published it [5]. In summary, the majority of earlier research suggests that linguistic and compositional characteristics of the material serve as the primary indicators for differentiating between fake news and legitimate news, which sums up the state-of-the-art in fake news detection systems. [9]. The linguistic and structural elements of news items are typically the focus of fake news detection systems. Still, they usually ignore contextual details like the news source's historical past or the sociopolitical context in which the news is being disseminated. These techniques, which were shown to have a low accuracy value, could not discern the context and semantic meaning of phrases taken from fake news [10].

The prevalent approach to detecting fake news, which is content-oriented, relies on natural language processing (NLP) to analyze text characteristics and identify fake news. By identifying linguistic patterns, such as word occurrences that are typical of satire, irony, sentiment, and topicality, natural language processing (NLP) tools evaluate news information. Fake news can also be distinguished by highlighting the textual qualities of the articles rather than the source or design elements in the headlines. This content-based method is predicated on the idea that real and fake news have different language and structural makeups. It offers operational guidance for a workable fake news detection system and suggests a hybrid algorithm combining a language approach and network cues [13]. Semantic aspects are also considered, along with using grammatical features through syntax parsing with Probabilistic Context-Free Grammar (PCFG) and identifying differences in words used between false and actual news items. discourse analysis findings and rhetorical structure were chosen as explanatory variables to ascertain whether or not the information was fabricated [14]. Term Frequency - Inverse Document Frequency (TF-IDF) is used in text analysis to characterize text characteristics.

III. PROPOSED METHOD

1) The following preparation procedures are used to reduce the vocabulary and eliminate terms that add no information in order to get the text ready for vectorization. (1)We opted for both punctuation and stop words removal, as well as word lemmatization. This decision was made to streamline the vocabulary and eliminate any linguistic inconsistencies.. Word vectorized representation technology, which was developed from the distributional hypothesis and enables words with similar meanings to have comparable representations, is generally referred to as word embeddings. An embedding model is used for generate vectorized representations of words in a dataset and ensure that they are in the same vector space.

2) To calculate the context similarity, representation calculates the separation between respective vectors in space. The data is sent into the input layer following preparation. the title of a news article and text are separated into smaller tokens by the input layer. A dictionary is also used to assign a distinct integer index to every token. To keep the input text at a constant length, padding is also used. Eventually, all of the text in the news article N is converted into numerical vectors, which are then indexed using a dictionary D..

3) LSTM was particularly better than RNN because it had a constant error carousel (CEC) module. By using a carefully thought-out gate structure, the Constant Error Carousel keeps error signals constant over time, guaranteeing that back propagated defects don't fade or worsen. The gate structure computes the internal value of CEC based on the current input values and the prior context as it switches to control data flow and memory.

The word embedding layer, the bidirectional Long Short-Term Memory (LSTM) layer, the fully connected hidden layer with 50 RELU (Rectified Linear Unit) units, the 50% dropout layer, and the softmax layer—which linearly transforms the output of the preceding layer to determine the probability scores of the labels—are the layers that make up this Recurrent Neural Network (RNN) model.

4) For the RNN classifier, we use an ADAM optimizer with a batch size of 32 and a learning rate of 0.0001. We evaluated the proposed methodology using the publicly available FakeNewsNet Dataset.Ultimately, an array of studies was carried out to exhibit the effectiveness of the suggested approach. This research's main contributions can be summed up as follows: utilizing deep hyper context to extract and assess content features at various orientations; An examination of advancements in hyperparameter optimization We demonstrate empirically that deep learning architecture specifically designed for the job of different class false news identification yields inferior results compared to ordinary machine learning algorithms trained with our suggested approach.

Advantages of proposed system: Extracting, enriching, and correlating highly valuable features semantically and contextually..Identifies semantic representation and hidden contextual information.Ensuring that embeddings are locally invariant within neighborhoods and enhance the discriminative effect.can improve a single short text semantic

IV. HARDWARE AND SOFTWARE REQUIREMENT

A. Backend Technologies:

1) Python: The system is built using the Python programming language, offering flexibility and a wide range of libraries for data analysis and machine learning.
2) NumPy: NumPy is utilized for numeric computing, providing powerful mathematical functions and array operations essential to processing.
3) Sci-learn (scikit-learn): Sci-learn is a Python library used for machine learning tasks, including classification, regression, clustering, and model evaluation.
4) Jupyter Notebook: Jupyter Notebook serves as the interactive computing environment for developing and presenting the system's code and analysis. It enables seamless integration of code, visualizations, and explanatory text, facilitating reproducible research and collaboration.

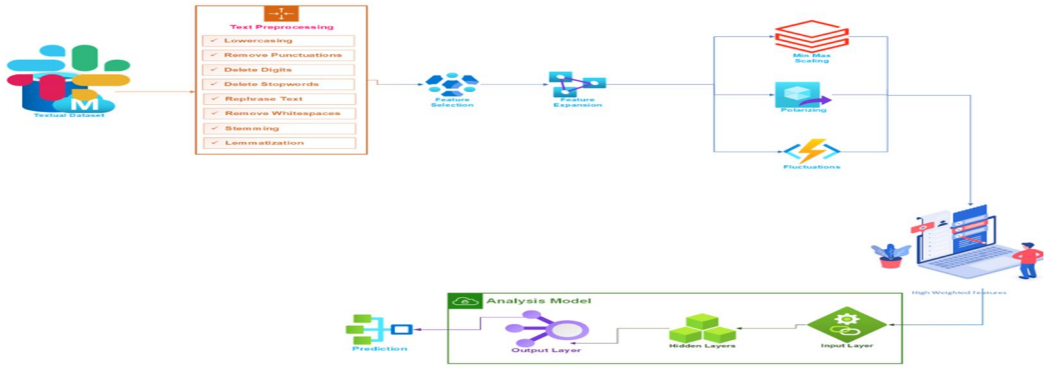
B. Frontend Technologies:

C. Web Technologies:

The frontend of the system utilizes web technologies for user interaction and visualization. The specific technologies employed may include:

1) HTML: HTML is used for structuring the content of web pages, providing a standardized markup language for creating web interfaces.
2) CSS: CSS (Cascading Style Sheets) used for creating web designs, allowing for customization of layout, colors, fonts, and other visual aspects.
3) JavaScript: JavaScript is employed for client-side scripting, enabling dynamic and interactive elements within web pages. Frameworks (e.g., React, Angular, Vue.js): Frontend frameworks may be used to facilitate the development of complex web applications, providing reusable components, state management, and routing capabilities.

V. ARCHITECTURE DIAGRAM



ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

VI. PROPOSED ALGORITHM

Proposed Algorithm: Probabilistic Analytical Learning Algorithm

The proposed algorithm, Probabilistic Analytical Learning Algorithm, leverages probabilistic and analytical techniques to facilitate effective cyberbullying detection and classification. By integrating principles from probability theory and analytical methods, the algorithm offers a robust framework for analyzing complex social media data and identifying cyberbullying instances.

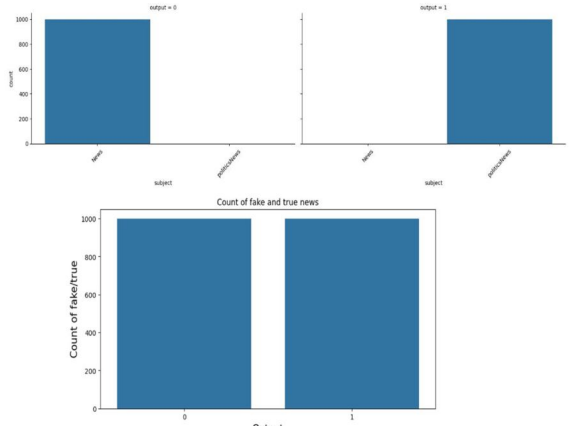
Advantages of Proposed Algorithm:

Ability to extract features and spotting Propaganda: The Probabilistic Analytical Learning Algorithm demonstrates resilience in scenarios where extracting news is made easier. By leveraging probabilistic reasoning and analytical capabilities, the algorithm can make informed decisions even with limited information.

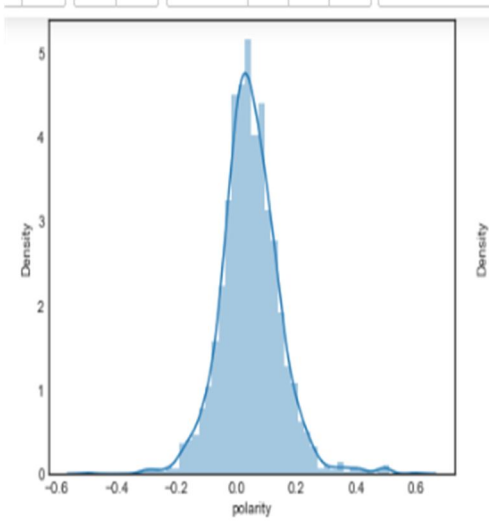
synthesizing content and contextual information in news stories to create a featured vector that improves the identification of fraudulent information. This directs the system to more accurately check for false and real.

The capability to capture the semantic relationships within text and retain extensive contextual information.. . By capitalizing on its combination it will make sure to cross check because of its processing capabilities and sophisticated analytical techniques, the algorithm can handle diverse inputs efficiently, ranging from data analysis and feature extraction to classification and decision-making.

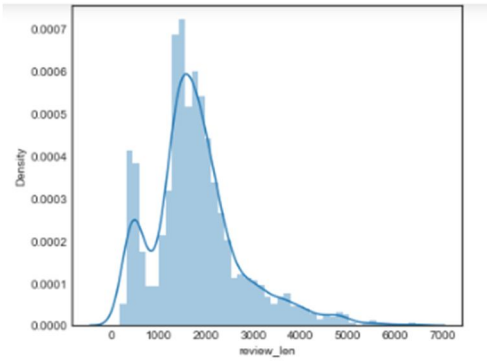
VII. RESULTS AND DISCUSSION



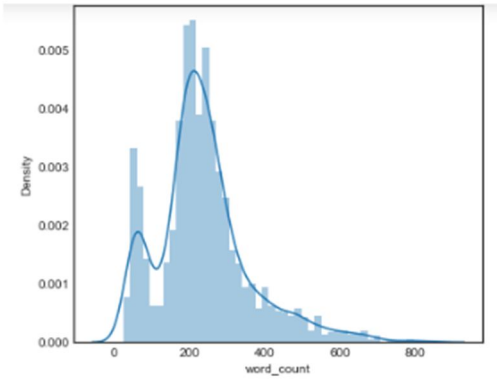
This graph determines the input of news between normal news and Political News and give the output to be 100% political news



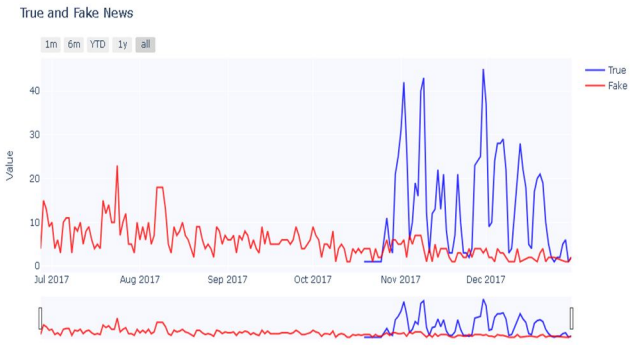
This Graph Represents the Polarity of the news



This Graph represents Review Length of the News



This Graph represents the variations in the word count of the news the is given for detection
The following graph represents the timeline of news articles that is produced between the year 2017 from JULY to DECEMBER, In which it reports that only less than 10 % of the articles produced are fake from the mid of OCTOBER to the end of DECEMBER and also shows the more than 40 % of the news produced are real



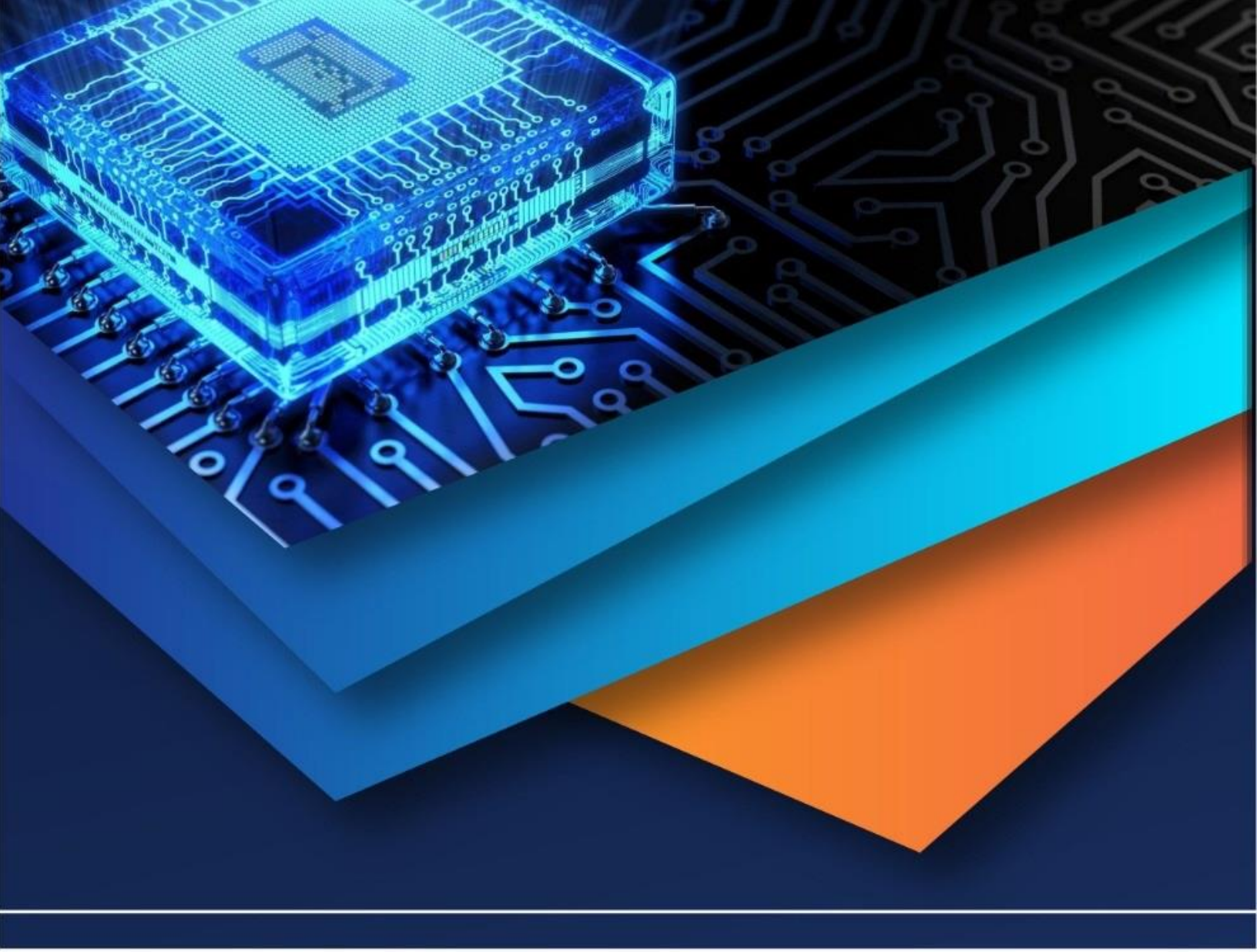
The outcomes demonstrate how well the suggested method works to identify false news. The OPCNN-false model demonstrated its supremacy in differentiating between real and false news stories by regularly outperforming other algorithms. The model excels in capturing intricate correlations among textual data by utilizing sophisticated approaches like word embeddings and hyperparameter optimization, together with CNN architecture.
Moreover, using feature extraction techniques as Glove word embeddings, TF-IDF, and N-gram improves the model's capacity to identify false news. All things considered, this highlights the potential of sophisticated machine learning to counteract disinformation, with the CNN-FAKE model providing a viable way to deal with this important problem.

VIII. CONCLUSION

This document outlines a structured approach for acquiring the required volume of data within a list generated by a data augmentation algorithm. The goal is to achieve balance among the minority labels in an imbalanced text dataset. A selection strategy is required to determine which items best enhance the dataset's performance for classification tasks, as most data augmentation models are capable of synthesizing an infinite number of items. Our selection process is based on choosing synthesized pieces that have the highest possible diversity and chance of belonging to their respective labels.

REFERENCES

[1] Minjung Park,Sangmi Chai Constructing a User-Centered Fake News Detection Model by Using Classification Algorithms in Machine Learning Techniques IEEE Access, 2023
[2] P. Bocij. "Fake news detection: Harassment in the Internet age and family. Greenwood Publishing to protect your Group. 2004." Noureddine Seddari,Abdelouahid Derhab,Mohamed Belaoued,Waleed Halboob,Jalal Al-Muhtadi,Abdelghani Bouras A Hybrid Linguistic and Knowledge-Based Analysis Approach for Fake News Detection on Social Media IEEE Access, 2022
[3] M. O. Lwin, B. Li, and R. P. Ang. ""Adolescents' Defense Strategies Against Online Harassment"." Journal of adolescence 35(1), 31-41, 2012.
[4] Hager Saleh,Abdullah Alharbi,Saeed Hamood Alsamhi OPCNN-FAKE: Optimized Convolutional Neural Network for Fake News Detection IEEE Access, 2021
[5] Myunghoon Kang,Jae Hyung Seo,Chanjun Park,Heuiseok Lim Utilization Strategy of User Engagements in Korean Fake News Detection IEEE Access, 2022
[6] Ying Guo,Wei Song A Temporal-and-Spatial Flow Based Multimodal Fake News Detection by Pooling and Attention Blocks IEEE Access, 2022
[7] Dhiren Rohera,Harshal Shethna,Keyur Patel,Urvish Thakker,Sudeep Tanwar,Rajesh Gupta,Wei-Chiang Hong,Ravi Sharma A Taxonomy of Fake News Classification Techniques: Survey and Implementation Aspects IEEE Access, 2022
[8] Hassan Ali,Muhammad Suleman Khan,Amer Alghadhban,Meshari Alazmi,Ahmad Alzamil,Khaled Al-Utaibi,Junaid Qadir All Your Fake Detector are Belong to Us: Evaluating Adversarial Robustness of Fake-News Detectors Under Black-Box Settings IEEE Access, 2021
[9] Wajiha Shahid,Yiran Li,Dakota Staples,Gulshan Amin,Saqib Hakak,Ali Ghorbani Are You a Cyborg, Bot or Human?â€”A Survey on Detecting Fake News SpreadersIEEE Access, 2022
[10] C. Wu, F. Wu, Y. Huang and X. Xie, Personalized news recommendation: Methods and challenges, ACM Trans. Inf. Syst., vol. 41, no. 1, pp. 1-50, Jan. 2023.
[11] X. Su, G. SperlÃ¬, V. Moscato, A. Picariello, C. Esposito and C. Choi, An edge intelligence empowered recommender system enabling cultural heritage applications, IEEE Trans. Ind. Informat., vol. 15, no. 7, pp. 4266-4275, Jul. 2019.
[12] F. Zhou, X. Xu, G. Trajcevski and K. Zhang, A survey of information cascade analysis: Models predictions and recent advances, arXiv:2005.11041, 2020.
[13] S. Gaillard, Z. A. OlÃ¡h, S. Venmans and M. Burke, Countering the cognitive linguistic and psychological underpinnings behind susceptibility to fake news: A review of current literature with special focus on the role of age and digital literacy, Front. Commun., vol. 6, Jul. 2021.
[14] E. C. Tandoc, Z. W. Lim and R. Ling, Defining â€œfake newsâ€: A typology of scholarly definitions, Digit. Journalism, vol. 6, no. 2, pp. 137-153, Feb. 2018.
[15] T. D. N. de Barcelos, L. N. Muniz, D. M. Dantas, D. F. Cotrim Junior, J. R. Cavalcante and E. Faerstein, AnÃ¡lise de fake news veiculadas durante a pandemia de COVID-19 no Brasil, Revista Panamericana de Salud PÃºblica, vol. 45, pp. e65, Jun. 2021.
[16] D. Carrion-Alvarez and P. X. Tijerina-Salina, Fake news in COVID-19: A perspective, Health Promotion Perspective., vol. 10, no. 4, pp. 290, 2020.
[17] X. Zhou and R. Zafarani, A survey of fake news: Fundamental theories detection methods and opportunities, ACM Comput. Surv., vol. 53, no. 5, pp. 1-40, 2020.
[18] K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang and Y. Liu, Combating fake news: A survey on identification and mitigation techniques, ACM Trans. Intell. Syst. Technol., vol. 10, no. 3, pp. 1-42, 2019.
[19] A. Bondielli and F. Marcelloni, A survey on fake news and rumor detection techniques, Inf. Sci., vol. 497, pp. 38-55, Sep. 2019
[20] F. D. Davis. "Perceived usefulness, perceived ease of use, and user acceptance of information technology." MIS Quarterly, 319-340, 1989.
[21] M. Norliza et al. "Women participation in business: A focus on franchising venture." Dept. Inf. Syst., Univ. Teknologi Malaysia, Malaysia, Tech. Rep. 104, Dec. 2006.
[22] W. M. Al-Rahmi et al. "Use of E-learning by University Students in Malaysian higher educational institutions: A case in Universiti Teknologi."
[23] P. Cohen, S. G. West, and L. S. Aiken. "Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences."



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY