Towards Mitigating Misinformation: A Structured Dataset of Fact-Checked Claims from News Media

Oam Bhanushali
K J Somaiya IoT, Mumbai, India
oam.b@somaiya.edu

Aditya Mer
K J Somaiya IoT, Mumbai, India
aditya.mer@somaiya.edu

Rishikesh Giridhar
K J Somaiya IoT, Mumbai, India
r.giridhar@somaiya.edu

Bhavormi Somaiya
K J Somaiya IoT, Mumbai, India
bhavormi.s@somaiya.edu

Arish Manasia
K J Somaiya IoT, Mumbai, India
arish.m@somaiya.edu

Shivam Singh
K J Somaiya IoT, Mumbai, India
shivam21@somaiya.edu

Neha Sharma
S4DS, Pune, India
nvsharma@rediffmail.com

Mrityunjoy Panday
S4DS, Pune, India
mrityunjoy@gmail.com

Milind Nemade
K J Somaiya IoT, Mumbai, India
mnemade@somaiya.edu

Sejal Shah
K J Somaiya IoT, Mumbai, India
sshah@somaiya.edu

G.S Mani
IEEE, Pune Section
gsmanihome@yahoo.com

2024 IEEE Region 10 Symposium (TENSYMP) | 979-8-3503-6486-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/TENSYMP61132.2024.10752132

***Abstract*—False information has become an unavoidable endeavor in the digital world, endangering public discourse and influencing people's decisions. Misinformation comes from a variety of sources and can travel quickly through online networks, drastically altering public perceptions, influencing political beliefs, and escalating audience prejudices. It is possible to come across several types of false information through text, picture, or video recordings. Misinformation has grown to be a significant problem in a world where memes are common, improved technologies are easily accessible, and mind-sharing is unrestrained. In order to address this widespread issue, researchers have gathered news information from many websites that provide truth-checking services, including India Today and ANI. Python libraries and modules, inclusive of BeautifulSoup and Selenium, were utilized to scrape those web sites and create the labeled dataset called "Factrix" indicating the veracity of the records (e.g., fake, true, half-true, mostly-false). Factrix dataset is intended to train models for the purpose of identifying false information through the use of textual records. These statistics can also be applied to pictorial analysis. Furthermore, these datasets enable trend analysis spanning from 2019 to 2024, providing insights into the evolution of incorrect information through the years.**

Keywords*—*Web Scraping, Data Labelling, Text Analysis, Trend Analysis, ColBert, Misinformation, Fake News

I. Introduction

In this Information Age, there has appeared a new problem that is worse than any before: misleading information. The internet has become a place full of deceit and lies—fake news, photoshopped pictures, and edited videos. The ”poison” of misinformation leaves institutions and the public polarized, and an absence of authentic debate ensues. The spreading of fake news has made advanced technology and identification crucial. New studies have opened the door to some promising alternatives that can be used effectively against this threat. Machine Learning, carefully patterned with labeled data, have the power to differentiate correctly between false and true information, doing so quite well. For example, machine learning algorithms were used to discover the truth versus falsehood, focusing on developing the most accurate model to determine whether the data is real or simulated [1]. However, digital threats involve more than just text. Hackers sometimes use tricked images and videos to support their lies. Image recognition is one solution that some medical researchers are even working on. For instance, the use of BERT and LSTM models is being analyzed for tackling non-factual information that involves both text and images [2]. The war against misinformation is a separate issue, but the relationship with the development of new detection technologies is interconnected, and strict control of the consumer's information depends upon it. A 2021 survey found that 30% of the survey participants in the United States supported the use of social media company-produced algorithms to detect false content [3]. The study also highlights the likelihood of driving a ”pause and chunk” thought process by encouraging the public to think twice about online information consumption. Additionally, 51% of the respondents favor the inclusion of social media platform

979-8-3503-6486-6/24/$31.00 ©2024 IEEE

in the process of setting standards for the programs. This is in line with behavioral science research on interventions meant to encourage users to choose to share accurate news. Traditional datasets are the very tools that power the development of state-of-the-art solutions for misinformation detection. These datasets, such as the MediaEval 2015 Image Verification Corpus [4], CASIA2GroundTruth [5], LIAR dataset [6], Politifact dataset [7], BuzzFeedNews Fake News Top 50 [8], MiSoVac [9], CoAID [10], COVID-19 Fake News Dataset [11], and Fake-NewsNet [12], include misinformation in areas such as political claims, COVID-19-related information, and image verification. The datasets offer researchers valuable materials for teaching and verifying detection models, facilitating a thorough analysis of the phenomenon. However, above mentioned datasets do not offer real-time news updates, as their accuracies reach 100% on machine learning models, representing an idealized scenario. The rest of the paper is organized in various section to enhance the reading of the paper. Section II reviews the spread of fake news and explores various detection methods, including machine learning techniques. Section III introduces the Factrix dataset, designed to aid in the detection of misinformation in the Indian context. Section IV outlines the implementation process, detailing data collection, labeling, and model development. Section V discusses additional considerations such as ethics and transparency in the detection process. Section VI presents the results, highlighting the performance metrics of the model trained on the Factrix dataset. Section VII concludes by emphasizing the effectiveness of the proposed approach in combating misinformation. Section VIII outlines future directions, suggesting the integration of advanced models and real time monitoring to enhance detection efforts.

II. Related Work

The unfold of fake news across many media structures, in particular social media, has turn out to be a vital hassle. Robust detection systems are essential because of the rapid spread of fraudulent information. Several approaches to this issue have been explored in the past, with a focus on machine learning techniques.

A. *Misinformation Detection in Different Media*

Researchers have attempted to find the way to identify fake content material on social networking sites. For instance, Jouhar et. al. have employed machine learning techniques to examine synthetic data, highlighting the creation of precise models capable of differentiating between fake and real data [1]. Furthermore, research has looked at situations in which heightened emotional states make false information more likely. Harbola et. al. used modern models like BERT and LSTM to enhance the identification of bogus information in emergency situations [2].

B. *Machine Learning for Misinformation Detection*

The use of ensemble strategies has shown promising results in enhancing the accuracy of erroneous fact identification. Through the combination of a couple of style-gaining knowledge of systems, Iftikhar et. al. confirmed the potential to generate extremely strong and reliable detection frameworks [12]. Unsupervised learning strategies offer an alternative approach via detecting anomalies in information that may imply false facts, while supervised mastering strategies, which rely upon categorized datasets, have been widely used [1][7]. Ashraf et. al. explored data augmentation techniques to expand the variety of training data to enhance model performance [11].

C. *Automated Misinformation Detection Systems*

A diverse array of approaches for automatically detecting false information has been proposed. Lexical-based techniques concentrate on linguistic analysis to identify patterns in sentiment, syntax, and word choice [5]. However, these methods may struggle with non-textual elements of misinformation. Deep learning models, such as LSTM and BERT, have shown improved performance in evaluations by capturing intricate relationships between words and context [2]. Multi-modal approaches which integrate both visual and textual components, are particularly effective in addressing misinformation that includes manipulated images alongside erroneous text [4].

D. *Understanding the Evolution of Misinformation*

A variety of methods for detecting false information have been proposed, with lexical approaches analyzing sentiment, syntax, and word choice. However, these techniques may face challenges with non-textual aspects of misinformation. Deep learning models like LSTM and BERT excel in capturing complex word relationships [2]. Multi-modal strategies which combine visual and textual elements, are particularly adept at tackling mis-information that involves altered images in conjunction with misleading text [4].

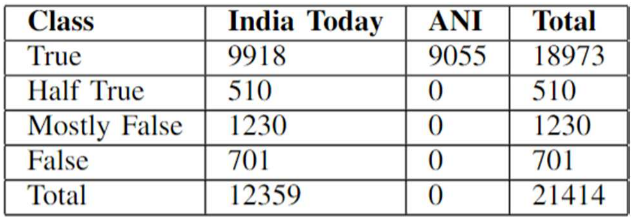
E. *Proposed Research*

This remark, which builds on contemporary research, objectives to assist ongoing initiatives to counter fake news. The objective is to enhance the fake information detection system by creating a comprehensive dataset, "Factrix" of data samples from credible sources such as India Today and ANI and applying machine learning techniques. This approach identifies false information by integrating text analysis, image recognition, and URL evaluation, enhancing defenses against digital misinformation using advanced machine learning.

III. Comprehensive Dataset for Analyzing Misinformation

This paper introduces a novel dataset "Factrix" constructed by scraping fact-checking services from prominent Indian news websites like India Today and ANI. Factrix dataset aims to support research on automated detection and classification of misinformation circulating in the Indian context as depicted in Table I. For a better understanding of the process refer to Fig 3.

Table I: Count of data in the dataset Factrix



Class	India Today	ANI	Total
True	9918	9055	18973
Half True	510	0	510
Mostly False	1230	0	1230
False	701	0	701
Total	12359	0	21414

A. Data Collection Methodology:

The data was collected through an automated web scraping process targeting the fact-checking sections of the aforementioned websites. The scraping script was designed to extract relevant information from the web page structure and store it in the defined JSON format as depicted in Fig 1. This ensures consistent data representation and facilitates further analysis.

B. Potential Applications:

Factrix dataset offers valuable resources for researchers exploring various aspects of misinformation detection:

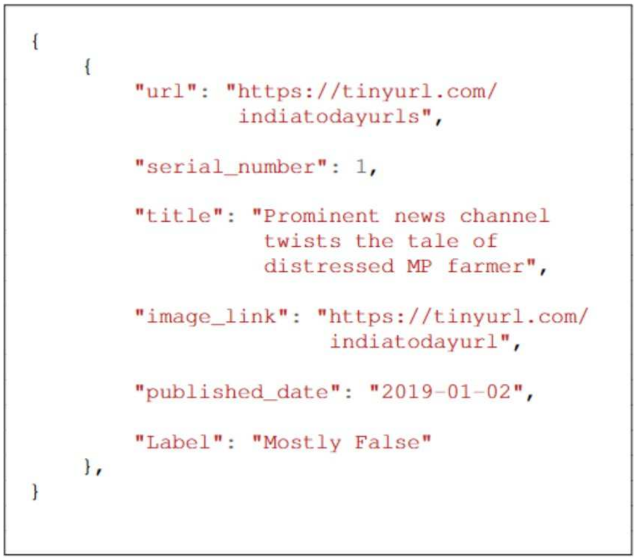
1) Machine Learning:: Model Training: Factrix dataset can be utilized to train machine learning models capable of automatically classifying claims as factual, misleading, or false. This can significantly improve the efficiency and scalability of misinformation detection efforts.

2) Misinformation Trend Analysis: Analyzing the types of claims most frequently fact-checked can provide insights into the prevalent themes and tactics used to spread misinformation in the Indian context.

3) Linguistic Analysis: Studying the language patterns used in both misinformation and factual content can help identify linguistic markers that differentiate between real and fake content. This knowledge can be incorporated into detection algorithms for improved accuracy.

4) Fact-Checking Strategy Evaluation: Factrix dataset can be used to evaluate the effectiveness of different fact-checking approaches employed by various websites. This can guide the development of more robust and comprehensive fact- checking strategies.

By providing a structured collection of fact-checked claims specific to the Indian context, Factrix dataset contributes significantly to the research efforts aimed at combating misinformation in the digital age. The diverse data elements and potential applications position this dataset as a valuable tool for researchers and developers working towards a more informed online environment.



```
{
    {
        "url": "https://tinyurl.com/
                indiatodayurls",

        "serial_number": 1,

        "title": "Prominent news channel
                  twists the tale of
                  distressed MP farmer",

        "image_link": "https://tinyurl.com/
                       indiatodayurl",

        "published_date": "2019-01-02",

        "Label": "Mostly False"
    },
}
```

Fig 1. Visual representation of data

IV. Implementation

Initially, the authors collected and annotated data by extracting information from fact-checking websites. They developed an elaborate labeling framework and archived the data in JSON format, as illustrated in Fig 1. Following this, they used text and URL analysis to build and train models, merging the findings for a comprehensive trend analysis. Finally, the authors assessed the model's performance and modified them to account for changing deception strategies. The entire process is illustrated in Fig 3.

- Programming Languages: Python
- Libraries/Modules: BeautifulSoup, Selenium, TensorFlow, PyTorch

The implementation is done sequentially in the following phases:

Phase 1: Data Collection & Labeling

A. Web Scraping:

- Develop Python scripts using BeautifulSoup and Selenium to extract data from identified fact-checking websites (India Today, ANI, etc.).
- Focus on extracting date, website URL, headline and any associated image/video URLs.
- Ensure robustness to handle website structure changes and potential anti-scraping measures.

B. Data Labeling:

- Develop a labeling schema based on the 4 categories: True, Half True, Mostly False, False.
- Consider using crowd sourcing platforms or employing human annotators for accurate labeling.
- Ensure consistency and inter-annotator agreement through clear guidelines and quality control measures.

C. Data Storage:

- Store the extracted and labeled data in JSON format as illustrated in Fig 1 for easy access and manipulation.
- Consider using a database for efficient storage and retrieval, especially for large datasets.

Phase 2: Model Development & Training

A. Textual Analysis:

- Experiment with different text pre-processing techniques (e.g., stop word removal, stemming) and feature engineering methods to optimize model accuracy.
- Using ColBert with Support Vector Machine (SVM) for training the model

B. URL Analysis:

- Develop methods to analyze the domain names, website reputation, and other relevant information associated with the URLs present in claims.
- Explore the use of blacklists or whitelists to identify potentially unreliable sources.

C. Trend Analysis

- Topic Identification and Fake News Distribution: Topic modelling is used to categorize articles by topic (e.g., Politics, Health). This helps identify topics where fake news is more prevalent, allowing for targeted countermeasures. As depicted in Fig 2.
- Trend Analysis of Fake News Creation: The analysis focuses on articles labeled as fake news. By analyzing the publication year of each fake news article, trends in fake news creation over time are visualized in Fig 2.

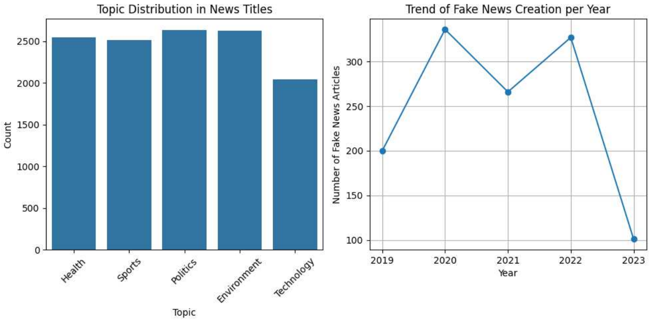


Fig 2. Topic Distribution in News Title and Trend of fake News Creation Per Year

D. Model Integration:

- Combine the insights from textual, and URL analysis into a comprehensive misinformation detection model.
- Investigate ensemble methods or other techniques to integrate the predictions from different models effectively.

Phase 3: Evaluation & Refinement

A. Model Evaluation:

- Utilize standard metrics such as accuracy, precision, recall, and F1 score to evaluate the performance of the developed models on a held-out test set.
- Conduct error analysis to identify areas for improvement and potential biases in the models.

B. Model Refinement:

- Based on the evaluation results, refine the models by adjusting hyperparameters, incorporating additional features, or exploring different model architectures.
- Continuously monitor and update the models to adapt to the evolving tactics of misinformation creators.

V. Additional Considerations:

- Ethical Implications: Address potential biases in the dataset and models, ensuring fairness and inclusivity in misinformation detection.
- Transparency and Explainability: Promote transparency by documenting the development process and providing clear explanations of model predictions.
- Data Privacy: Ensure responsible data collection and handling practices, respecting user privacy and complying with relevant regulations.

VI. Results

The following classification report presents the performance metrics of the model trained using ColBERT with SVM for misinformation detection within Factrix dataset. The metrics include precision, recall, and F1-score for each class (False, Half True, Mostly False, and True), as well as the support for each class. These metrics provide insight into how well the model distinguishes between different levels of truthfulness in the dataset. Additionally, the authors report the overall accuracy and the Matthews Correlation Coefficient (MCC) of the model. While accuracy is a common performance metric, it can be misleading in the context of imbalanced datasets, such as the one used in this study. Therefore, MCC is also provided as it offers a more balanced evaluation by considering both true and false positives and negatives.

Table II: Tabular representation of classification metrics

Class	**Precision**	**Recall**	**F1-Score**	**Support**
False	0.56	0.16	0.24	128
Half True	1.00	0.01	0.02	93
Mostly False	0.45	0.36	0.40	219
True	0.94	1.00	0.97	3843

- **Accuracy:** 0.92

- **Matthews Correlation Coefficient (MCC):** 0.451

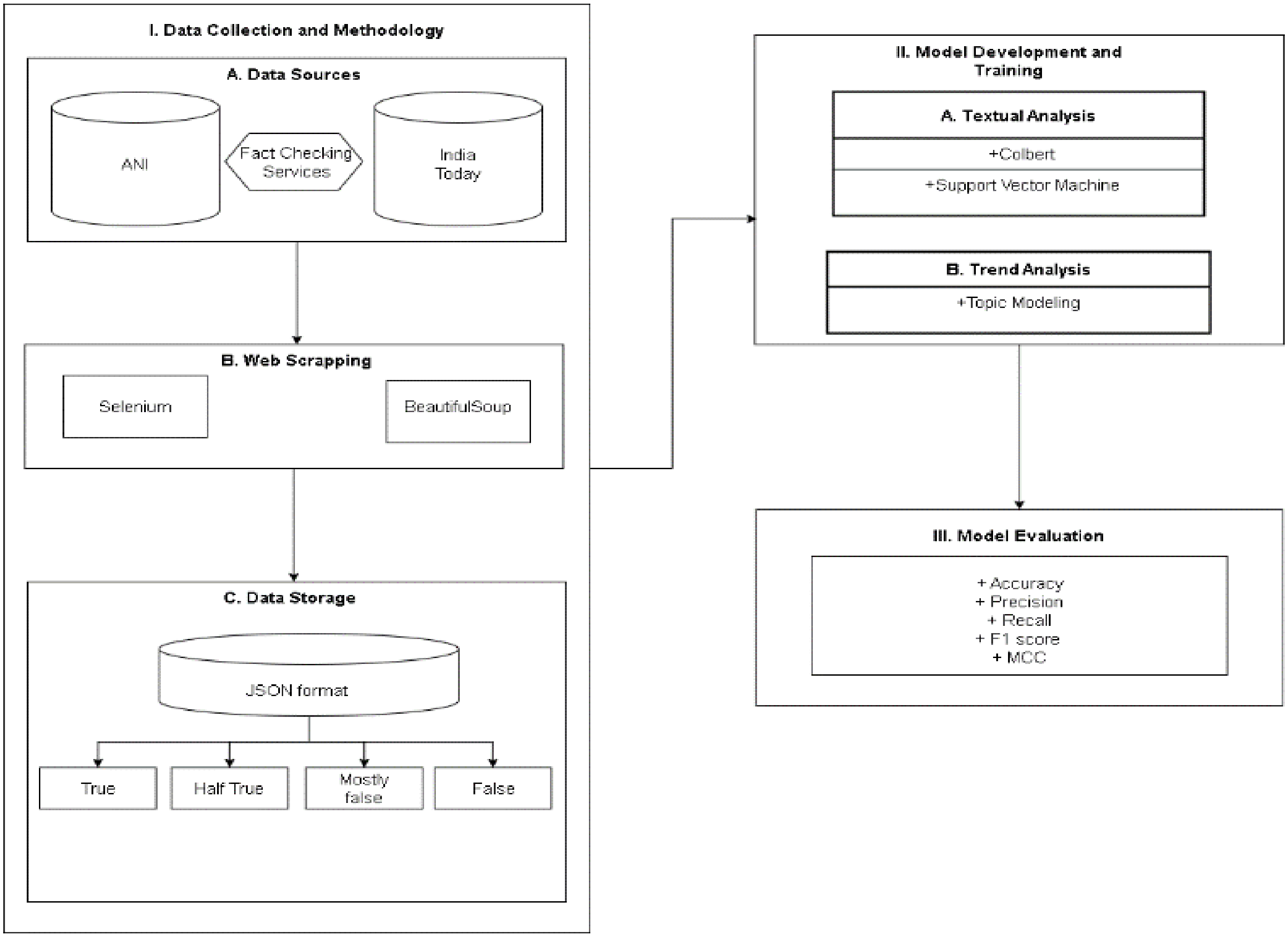


Fig 3. Process diagram

VII. Conclusion

In conclusion, the research gives a multi-faceted technique to fight misinformation in the digital age. By leveraging machine learning algorithms and web scraping strategies, the authors have advanced a framework for figuring out and mitigating the unfold of misinformation. Through the gathering and analysis of records from reliable sources, the authors have created categorized dataset named “Factrix” to train models for detecting misinformation. The classification file generated the usage of ColBERT with SVM highlights the overall performance metrics of the model. The precision, recall, and F1-scores for the numerous instructions (False, Half True, Mostly False, and True) suggest the model’s capability to differentiate amongst distinct levels of truthfulness in the dataset. Additionally, the general accuracy of the model is mentioned as 0.92, even as the Matthews Correlation Coefficient (MCC) is calculated to be 0.451.

Although accuracy is a common overall performance metric, it may be misleading because of the biased nature of the dataset. Hence, MCC is likewise supplied because it gives a more balanced assessment via considering each true and fake positives and negatives. As technologies continue to evolve, the methods proposed by the authors will adapt to address growing challenges. By promoting collaboration, education, and the integration of advanced models, the authors aim to foster a more knowledgeable and resilient online environment for all users.

VIII. Future Scope

Incorporating modern fashions collectively with LLama (Large Language Model Meta AI) and LLaVA (Large Language and Visual Assistance) for Image assessment can In- crease the accuracy of incorrect information detection. These models leverage multi-version reading strategies to at the same time examine textual and visible content, permitting more strong identification of fake pictures and assessing the Image’s credibility. Developing structures for actual-time monitoring of misinformation developments can provide well timed insights into rising threats. Integrating natural language processing and image evaluation algorithms into computerized monitoring equipment can assist identify and counteract false information because it spreads across on line platforms. Educating customers about the dangers of incorrect information and imparting equipment to verify the credibility of information can empower customers to make more informed choices on-line. Future studies may want to attention on growing consumer-friendly interfaces and academic campaigns aimed toward selling crucial wondering and digital literacy. Collaborating with international organizations, fact-checking initiatives, and social media platforms can amplify the impact of misinformation detection efforts. Sharing datasets, best practices, and detection algorithms across borders can help address the global challenge of online misinformation more effectively. To combat misinformation, the authors need contemporary technology, real-time monitoring, improved user awareness, and worldwide collaboration. Tools such as LLama and LLaVA analyze textual content and imagery to detect bogus news. Promoting critical thinking and preventing misinformation from spreading is essential. Working together globally increases the effectiveness in combating fake news.

ACKNOWLEDGEMENT

The authors express sincere gratitude to Society for Data Science for providing the problem statement and mentoring closely to complete the project successfully. All authors have contributed equally to this work.

REFERENCES

[1] Jouhar, J., Pratap, A., Tijo, N., Mony, M. (2024). Fake News Detection using Python and Machine Learning. Procedia Computer Science, 233, 763–771. https://doi.org/10.1016/j.procs.2024.03.265

[2] Harbola, A., Manchanda, M., Negi, D. (2023). Misinformation classification using LSTM and BERT model. Presented at the 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA). https://doi.org/10.1109/icidca56705.2023.10100054

[3] Pew Research Center (2021). News Consumption Across Social Media in 2021.

[4] Michael, M., Das Bhattacharjee, S., Yuan, J. (2023). Self-Supervised Distilled Learning for Multi-modal Misinformation Identification. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 284-293. https://doi.org/10.1109/WACV56688.2023.00284

[5] Ramakrishnan, K., Balakrishnan, V. (2022). Health Misinformation in the Covid-19 Era - Detecting Misinformation on Bilingual Corpora using Lexical Features. 2022 International Conference on Electronics, Computing and Communication Technologies (CONECCT), 1-5. https://doi.org/10.1109/CONECCT55909.2022.9988481

[6] (2022). Misinformation in Social Media Platforms and Web Articles: a Dataset to Infer User Stance. 2022 IEEE International Conference on Semantic Computing (ICSC), 51-58. https://doi.org/10.1109/icsc52841.2022.00051

[7] Imbwaga, J.L., Chittaragi, N.B., Koolagudi, S.G. (2022). Fake News Detection Using Machine Learning Algorithms. Proceedings of the 5th International Conference on Artificial Intelligence and Pattern Recognition, 1-6. https://doi.org/10.1145/3549206.3549256

[8] Diaz Ruiz, C., Nilsson, T.K.H. (2022). Disinformation and Echo Chambers: How Disinformation Circulates on Social Media Through Identity-Driven Controversies. Journal of Public Policy & Marketing, 41(2), 169-182. https://doi.org/10.1177/07439156221103852

[9] Di Domenico, G., Nunan, D., Pitardi, V. (2022). Marketplaces of Misinformation: A Study of How Vaccine Misinformation Is Legitimized on Social Media. Journal of Public Policy & Marketing, 41(2), 183-196. https://doi.org/10.1177/07439156221103860

[10] Olan, F., Jayawickrama, U., Arakpogun, E.O., Suklan, J., Liu, S. (2022). Fake news on Social Media: the Impact on Society. Information Systems Frontiers, 24, 1773-1794. https://doi.org/10.1007/s10796-022-10242-z

[11] Butt, A.N., Sidorov, G., Gelbukh, A. (2021). CIC at CheckThat! 2021: Fake News detection Using Machine Learning And Data Augmentation. CLEF 2021 Working Notes, 446-454. http://ceur-ws.org/Vol-2936/paper-34.pdf

[12] Ahmad, I., Yousaf, M., Yousaf, S., Ahmad, M.O. (2020). Fake News Detection Using Machine Learning Ensemble Methods. Complexity, 2020, 8885861. https://doi.org/10.1155/2020/8885861

[13] (2023). Explainable Misinformation Detection across Multiple Social Media Platforms. IEEE Access, 11, 23634-23646. https://doi.org/10.1109/ACCESS.2023.3251892

[14] Khan, E.M., Rath, B., Vraga, E.K., Srivastava, J. (2023). Behavioral Forensics in Social Networks: Identifying Misinformation, Disinformation and Refutation Spreaders Using Machine Learning. arXiv preprint arXiv:2305.00957. https://doi.org/10.48550/arXiv.2305.00957

[15] Kydd, M., Shepherd, L.A. (2023). Deep Breath: A Machine Learning Browser Extension to Tackle Online Misinformation. arXiv preprint arXiv:2301.03301. https://doi.org/10.48550/arXiv.2301.03301

[16] Ling, W., & Shi, W. (2023). Efficacy of Educational Misinformation Games. arXiv.org. https://doi.org/10.48550/arXiv.2305.09429

[17] Singh, B., & Sharma, D. K. (2021). Predicting image credibility in fake news over social media using a multi-modal approach. Neural Computing and Applications, 1-15. https://doi.org/10.1007/S00521-021-06086-4

[18] Yoon, S., Park, K., Lee, M., Kim, T., Cha, M., & Jung, K. (2021). Learning to Detect Incongruence in News Headline and Body Text via a Graph Neural Network. IEEE Access. https://doi.org/10.1109/ACCESS.2021.3062029

[19] Bojjireddy, S., Chun, S. A., & Geller, J. (2021). Machine Learning Approach to Detect Fake News, Misinformation in COVID-19 Pandemic. https://doi.org/10.1145/3463677.3463762

[20] Choudhary, M., Jha, S., Saxena, D., & Singh, A. K. (2021). A Review of Fake News Detection Methods using Machine Learning. https://doi.org/10.1109/INCET51464.2021.9456299

[21] Luo, H., Cai, M., & Cui, Y. (2021). Spread of Misinformation in Social Networks: Analysis Based on Weibo Tweets (C. Gan, Ed.). Hindawi Limited. https://doi.org/10.1155/2021/7999760

[22] Shi, Z., Liu, X., & Srinivasan, K. (2021). EXPRESS: Hype News Diffusion and Risk of Misinformation: The Oz Effect in Healthcare. Journal of Marketing Research. https://doi.org/10.1177/00222437211044472

[23] Ghai, S. (2020). Machine Learning to Limit the Spread of Misinformation. International Journal of Engineering Development and Research (IJEDR), 8(3), 248-252. http://www.ijedr.org/papers/IJEDR2003040.pdf

[24] Tran, T., Rad, P., Valecha, R., & Rao, H. R. (2020). Misinformation in Crises: A Conceptual Framework for Examining Human-Machine Interactions. https://doi.org/10.1109/AI4G50087.2020.9311010

[25] Allcott, H., Gentzkow, M., & Yu, C. (2019). Trends in the diffusion of misinformation on social media. Research Politics. https://doi.org/10.1177/2053168019848554

[26] Zucker, A. (2019). Using critical thinking to counter misinformation. Science Scope, 42(8), 6–9. https://www.jstor.org/stable/26898998

[27] Koohang, A., & Weiss, E. (2003). Misinformation: Toward Creating a Prevention Framework. https://doi.org/10.28945/2603