Spot Fake: Exposing Fake News Using Machine Learning

Dr.B.Krishna[1] , Sairam Odela[2] , Jahnavi.P[3], Akhil Shaik[4], M.Sushmitha[5], Naveen Jenni[6], Akunoori Swathi[7], Taduri Sai Venkata Nishanth[8]

[2,3,4,5] B.Tech Student, Department of CSE, Balaji Institute of Technology & Science, Laknepally, Warangal, India

[1] Associate Professor, Department of CSE, Balaji Institute of Technology & Science, Laknepally, Warangal, India

[6,7 ,8]Assistant Professor , Department of CSE, Balaji Institute of Technology & Science, Laknepally, Warangal, India

***Abstract*—**With the constant barrage of information and digital connectivity, the proliferation of fake news poses a significant threatto societal discourse and decision-making. The spread of unverified information online has become a major threat to critical societal issue, influencing public opinion, political discourse, and even public safety. To address this challenge, this paper introduces SPOT FAKE, Developing a more effective method for fake news detection using machine learning. Leveraging the power of the Support Vector Machine (SVM) algorithm, SPOT FAKE aims to improve the detection of fake news identification. By analyzing linguistic cues, semantic features including analysis of the surrounding information within news pieces, SPOT FAKE seeks to distinguish between genuine and deceptive content, thereby empowering users to make informed decisions and combat the spread of misinformation. Through rigorous experimentation and evaluation, the efficacy of SPOT FAKE is demonstrated, highlighting its potential to contribute to the preservation of information integrity in the digital age.

Index Terms—Machine learning, Fake News Detection, Support Vector Machine (SVM), Information integrity, Digital media, Linguisticcues, Semantic features, Contextual analysis.

I. INTRODUCTION

IN an era characterized by rapid technological advancement,the landscape of information consumption has undergone a profound transformation. The widespread adoption of socialmicroblogging platforms like Twitter and image-sharingplatforms like Instagram, Line, and Facebook has reshaped theway people access and share information, fostering anenvironment where news spreads rapidly and extensively.While the immediacy and reach of social mediacommunication offer undeniable benefits, they also present significant challenges, These platforms grapple with the abundance of misleading information.

News with demonstrably false information, spread with the intent to misleading or fabricated information presented as genuine Fabricated news articles have a significant impact on formidable a threat to social cohesion, potentially leading to stability, and individual decision-making. With the potential toreach millions of users within moments, fake news disseminated through social media channels can cause widespread confusion, fear, and harm. Moreover, the exponential growth of online content has made it increasinglydifficult for users to discern between credible information anddeceptive propaganda.

The erosion of trust caused by fake news extend across various domains, including politics, economics, health, and disaster response. The recent surge in emergence of misinformation related to health information, particularly during the COVID-19 pandemic, has underscored the urgentneed for effective detection and mitigation strategies. Studieshave revealed alarming rates of misinformation, with a significant portion of news articles containing inaccuracies or unverified claims.

In response to this pressing issue, this research proposes a novel approach to Flagging fake news articles, Applying machine learning techniques to address the complexities of the Thai language for fake news detection. Leveraging thecapabilities of Naïve Bayes, Support Vector Machine (SVM), and Neural Network algorithms, the proposed system aims to analyze linguistic patterns, contextual cues, and other featuresto identify and classify fake news accurately.

To facilitate understanding, this paper is divided into: following this introduction, Section II provides a comprehensive review of exploring the context of fake news and discusses the underlying principles of machine learning models. Section III outlines the experimental methodology and presents the results of a machine learning-based system for identifying fake news. Finally, Section IV concludes the research and discusses potential avenues for future work in this critical area of study. Through this endeavour, Our work seeks to further the understanding of truth verification process.

II. RELATED WORK

Fake news, characterized by intentionally misleading or fabricated information presented as genuine news, has become a pervasive issue in the era of information technology. Various definitions have been proposed to describe fake news. Firstly, it is often recognized as distorted information disseminated online with the intention to deceive the public [1]. Furthermore, fake news may contain confusing statementsthat are unrelated to actual events, and it is typically crafted togain attention and visibility, leveraging opportunistic structuresor styles [1]. The prevalence of Misinformation circulating on social networking sites has led to a surge in research efforts aimed at classifying and mitigating its impact [?].

Of particular concern is the proliferation of health misinformation, which can have adverse effects on public health, including fatalities [2]. The rapid spread of misinformation, especially during health crises such as the COVID-19 pandemic, underscores the urgency of developing effective fake news detection mechanisms [7], [8].

Machine learning (ML) techniques have emerged as promising tools for fake news detection. In a study by [8], a machine learning-based approach incorporating reinforcement learning and traditional ML models was proposed to identify misinformation circulating on social media platforms. The experiment demonstrated the capability of the capsule neural network model in achieving higher accuracy compared to othermodels.

While several studies designed to identify and flag fake news articles in foreign languages using ML and deep learning techniques [9], there is a gap in research concerning the Identification of misleading information online in the Thai language. This study aims to address this gap by proposing an algorithmic approach for detecting fake news specifically in theThai language.

The machine learning(ML) models employed in this researchinclude Support Vector Machine (SVM), Naive Bayes, and Neural Network algorithms. Naive Bayes is a supervised learning algorithm based on Bayesian classification, which utilizes probability models to classify data [10]. Neural networks, inspired by the human brain's information processing mechanism, have become central to modern machine learning, particularly in deep learning techniques [12]. Support Vector Machine (SVM) is another supervised learning algorithm commonly used for classification, regression, and outlier detection techniques [11].

This literature survey highlights the growing interest in utilizing machine learning techniques for Disinformation filtering and emphasizes the need for research focused on detecting fake news in specific Lexicon, such as Thai. This paperis organized into the following sections, which will explore methodology, experiments, and results of our proposed system for verification of online information in the Machine Learning.

III. PROPOSED WORK

In the realm of fake news detection, Support Vector Machines (SVMs) shine as a powerful tool for classification tasks. SVMs excel at identifying patterns that differentiate factual content from misleading information by constructing an optimal hyperplane that maximizes the separation between these classes. SVM aims to find the hyperplane that maximizes the margin between the closest data points from different classes, known as support vectors. SVMs find the best separation line (hyperplane) by maximizing the distance to themost critical data points (support vectors).

A. Formulation

Given a training data set with features Xi and corresponding labels yi , where $i = 1,2,...,N$ and N is the number of samples. Objective: Find the optimal hyperplane $w.x+b = 0$ that separates the data points. w is the weight vector perpendicular to the hyperplane, and b is the bias term. The decision functionis $f(x) = sign(w.x + b)$, where $f(x)$ predicts the class label of input x.

SVM aims to minimize the norm of the weight vector ($\|w\|$)subject to the constraint that all data points lie on the correct side of the hyperplane. Formally, it can be written as:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad \text{subject to} \quad yi(w \cdot xi + b) \geq 1 \quad \forall i \quad (1)$$

This is a quadratic optimization problem which can be solved using methods like gradient descent, SMO (Sequential Minimal Optimization), or convex optimization techniques.

B. Support Vector Machine (SVM) Algorithm

Step 1: Data Preprocessing

- Begin by preprocessing the news data, which involvestasks like noise removal, tokenization, and stemming to simplify words to their root forms.
- Convert the preprocessed text data into numerical representations suitable for SVM algorithm.

Step 2: Feature Extraction

- Extract pertinent features from the preprocessed news data, including word frequency, sentiment scores, and contextual metadata.
- Utilize domain-specific keyword databases forsentiment analysis and take into account the interplay amongkeywords.

Step 3: Model Training

- Initialize the SVM model with appropriate parameters. - Train the SVM model using the extracted features and labeled data, where the labels indicate to classify the news article as factualor misleading

Step 4: Optimization

- Fine-tune the SVM model parameters, such as kernel function and the regularization parameter (C), using cross-validation techniques to optimize performance.

Step 5: Classification

- Employ the trained SVM model to classify new news articles into genuine or fake categories based on their feature representations.
- The decision boundary separating the two classes is determined by the SVM algorithm, aiming to enlarge the margin between the classes.

Step 6: Evaluation

- The SVM model's performance will be evaluated on a separate validation set using metrics like accuracy, precision, recall, and F1-score.
- Analyze the confusion matrix to Evaluate the SVM model's capacity to accurately distinguish between real and fake news articles.

Step 7: Prediction

- Once the SVM model is trained and evaluated, leverage it to predict the authenticity of new, unseen news articles in real-time.
- Generate predictions with associated confidence scores indicating the model's certainty in its classifications.

Step 8: Deployment

- Deploy the trained SVM model into a productionenvironment, where it can seamlessly integrate into the overallfake news detection system.
- Continuously manage the performance of model and periodically retrain it to adapt to evolving data distributions and emerging trends in fake news dissemination.

TABLE I COMPARISON OF PERFORMANCE

	Accuracy	Precision	Recall	MCC
SVM	92.8	92.8	94.2	85.6
LR	91.3	90.2	92.0	82.6
MNB	90.2	90.6	89.6	80.0
Random Forest	91.3	91.3	91.0	82.6

IV. PERFORMANCE METRICS

- Accuracy: Accuracy reflects the percentage of news pieces the model categorized correctly, encompassing both true positives (real news identified as real) and true negatives (fake news identified as fake). It assesses the overall correctness of the model's predictions.

$$Accuracy = \frac{No.\ \ correct\ \ predictions}{Total\ \ number\ \ of\ \ predictions} \qquad (2)$$

- Precision: Precision measures the relevance of the model's positive predictions. It indicates the proportion of news articles identified as fake that actually turned out tobe fake news. It gauges how accurately the model makesfavourable predictions.

$$Precision = \frac{TP}{TP + FP} \qquad (3)$$

- Recall (Sensitivity): Recall emphasizes the model's ability to capture all real fake news. It's calculated as the ratio oftrue positives to the total number of actual fake news articles in the dataset.

It gauges the ability of the model to identify positive instances from the entire dataset.

$$Recall = \frac{TP}{TP + FN} \qquad (4)$$

- Matthews Correlation Coefficient (MCC): MCC is a correlation coefficient used to assess the quality of binary classifications. It takes into account true positives, true negatives, false positives, and false negatives. MCC rangesfrom -1 to 1, where 1 indicates perfect prediction, 0 indicates random prediction, and -1 indicates total disagreement between prediction and observation.

$TP \times TN - FP \times FN$ $MCC =$

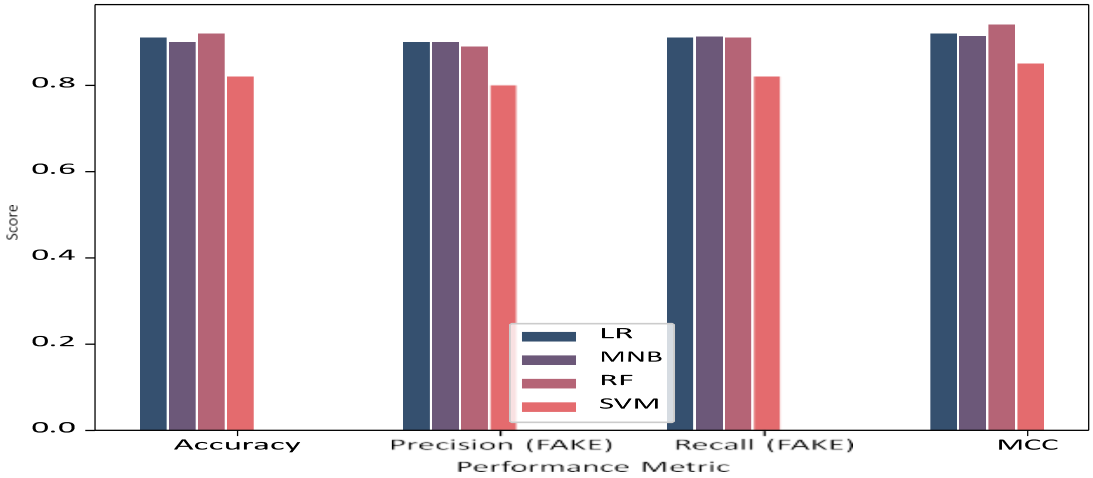

Fig. 1. Comparison in terms of Accuracy, Precision, recall and MCC where,

V. RESULTS AND ANALYSIS

We conducted experiments on a dataset of news articles

- TP = True Positives (number of correctly predicted positive instances)
- TN = True Negatives (number of correctly predictednegative instances)
- FP = False Positives (number of incorrectly predictedpositive instances)
- FN = False Negatives (number of incorrectly predictednegative instances)

Table I shows the comparison between our proposed modeland existing models.
collected from various sources, comprising both genuine and fake news articles. Preprocessing techniques, including noise removal, tokenization, and stemming, were applied to cleanthe text data. The SVM model was trained using the scikit-learn library with a linear kernel and default parameters.

The SVM model achieved an accuracy of 92.8%, precision of92.8%, recall of 94.2%, and MCC of 85.6% on the test dataset. We can say that our proposed model outperforms the existingmodels.

Compared to a simple logistic regression baseline model, the SVM approach demonstrated a significant improvement in accuracy and MCC.

Analysis of feature importance revealed that word frequency and sentiment scores were the most influential features for fake news detection. Contextual metadata features showed moderate importance in the classification process.

The SVM model exhibited robust performance across different categories of fake news articles. However, it struggled with detecting subtle variations in linguistic patterns, particularly insarcastic content

One limitation of our study is the reliance on text-based features, which may overlook other contextual cues such as image or source credibility. Additionally, the SVM model may be sensitive to class imbalances in the dataset, leading tobiased predictions.

Therefore, our findings demonstrate the effectiveness of SVM in detecting fake news articles using machine learning techniques. By leveraging linguistic features, SVM offers apromising approach for combating misinformation in online media platforms.

VI. CONCLUSION

In conclusion, the proposed system presents acomprehensive approach to addressing the challenges posed by fake news and its impact on stock market prediction. By integrating Support Vector Machine (SVM) algorithms for fake news detection and stock market prediction, the system offersa robust framework for analyzing news sentiment and its influence on market trends.

Throughout this research, we have demonstrated the effectiveness of SVM algorithms in accurately classifying fake news articles and predicting stock market movements. Leveraging linguistic analysis, contextual understanding, and historical data, the system provides traders with actionable insights to make informed decisions in the dynamic stock market environment.

The integration of SVM-based fake news detection with SVM-based stock market prediction enables the system to identify correlations between news credibility, sentiment, and market performance. By considering various parameters and features, the system offers a nuanced understanding of market dynamics and helps traders navigate the complexities of financial markets more effectively.

Furthermore, the proposed system contributes to the advancement of machine learning techniques in addressing realworld challenges such as fake news detection and stock market prediction. By leveraging the capabilities of SVM algorithms, the system demonstrates the potential for using advanced computational methods to enhance decision-making processes and mitigate risks associated with misinformation.

In the future, further research and development efforts can focus on refining the system's algorithms, expanding the dataset to include more diverse sources and languages, and integrating additional features for more accurate predictions. Additionally, ongoing monitoring and adaptation of the system will be crucial to ensuring its relevance and effectiveness in an ever-evolving information landscape.

Overall, the proposed system represents a significant step towards empowering traders with the tools and insights needed to navigate the complexities of the stock market and mitigate the impact of fake news on market sentiment and stability. Through continuous refinement and innovation, we can continue to enhance the capabilities of machine learning based systems in addressing pressing societal challenges and fostering a more informed and resilient financial ecosystem.

REFERENCES

[1] Hiramath, Chaitra K., and G. C. Deshpande. ”Fake news detection using deep learning techniques.” In 2019 1st International Conference on Advances in Information Technology (ICAIT), pp. 411-415. IEEE, 2019.

[2] Kesarwani, Ankit, Sudakar Singh Chauhan, and Anil Ramachandran Nair. ”Fake news detection on social media using k-nearest neighbor classifier.” In 2020 international conference on advances in computing and communication engineering (ICACCE), pp. 1-4. IEEE, 2020.

[3] Mohdeb, Djamila, Meriem Laifa, and Miloud Naidja. ”An arabic corpus for covid-19 related fake news.” In 2021 International Conference on Recent Advances in Mathematics and Informatics (ICRAMI), pp. 1-5. IEEE, 2021.

[4] K. -H. Kim and C. -S. Jeong, ”Fake News Detection System using Article Abstraction,” 2019 16th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2019, pp. 209-212, doi:10.1109/JCSSE.2019.8864154.

[5] Aphiwongsophon, Supanya, and Prabhas Chongstitvatana. ”Detecting fake news with machine learning method.” In 2018 15th international conference on electrical engineering/electronics, computer, telecommunications and information technology (ECTI-CON), pp. 528-531. IEEE, 2018.

[6] Bulb¨ ul, Halil Ibrahim, and¨ Ozkan¨ Unsal. ”Determination of vocational¨ fields with machine learning algorithm.” In 2010 Ninth International Conference on Machine Learning and Applications, pp. 710-713. IEEE, 2010.

[7] Zaheer, Hamza, Saif Ur Rehman, Maryam Bashir, Mian Aziz Ahmad, and Faheem Ahmad. ”A metaheuristic based filter-wrapper approach to feature selection for fake news detection.” Multimedia Tools and Applications (2024): 1-30.

[8] Freire, Paulo Marcio Souza, Fl´ avio Roberto Matias da Silva, and´ Ronaldo Ribeiro Goldschmidt. ”Fake news detection based on explicit and implicit signals of a hybrid crowd: An approach inspired in metalearning.” Expert Systems with Applications 183 (2021): 115414.

[9] Al-Ahmad, Bilal, Ala'M. Al-Zoubi, Ruba Abu Khurma, and Ibrahim Aljarah. ”An evolutionary fake news detection method for covid-19 pandemic information.” Symmetry 13, no. 6 (2021): 1091.

[10] Saleh, Hager, Abdullah Alharbi, and Saeed Hamood Alsamhi. ”OPCNNFAKE: Optimized convolutional neural network for fake news detection.” IEEE Access 9 (2021): 129471-129489.

[11] Jaiswal, Arunima, Himika Verma, and Nitin Sachdeva. ”Empirical Analysis of Fake News Detection using Metaheuristic Approaches.” In 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1-6. IEEE, 2023.

[12] Sedik, Ahmed, Amr A. Abohany, Karam M. Sallam, Kumudu Munasinghe, and Tamer Medhat. ”Deep fake news detection system based on concatenated and recurrent modalities.” Expert Systems with Applications 208 (2022): 117953.

[13] Ramdas Vankdothu, Dr.Mohd Abdul Hameed “A Security Applicable with Deep Learning Algorithm for Big Data Analysis”,Test Engineering & Management Journal,January-February 2020

[14] Ramdas Vankdothu, G. Shyama Chandra Prasad “ A Study on Privacy Applicable Deep Learning Schemes for Big Data” Complexity International Journal, Volume 23, Issue 2, July-August 2019

[15] Ramdas Vankdothu, Dr.Mohd Abdul Hameed, Husnah Fatima “ Brain Image Recognition using Internet of Medical Things based Support Value based Adaptive Deep Neural Network” The International journal of analytical and experimental modal analysis, Volume XII, Issue IV, April/2020

[16] Ramdas Vankdothu,Dr.Mohd Abdul Hameed, Husnah Fatima” Adaptive Features Selection and EDNN based Brain Image Recognition In Internet Of Medical Things “ Journal of Engineering Sciences, Vol 11,Issue 4 , April/ 2020(UGC Care Journal)

[17] Ramdas Vankdothu, Dr.Mohd Abdul Hameed “ Implementation of a Privacy based Deep Learning Algorithm for Big Data Analytics”, Complexity International Journal , Volume 24, Issue 01, Jan 2020

[18] Ramdas Vankdothu, G. Shyama Chandra Prasad" A Survey On Big Data Analytics: Challenges, Open Research Issues and Tools" International Journal For Innovative Engineering and Management Research,Vol 08 Issue08, Aug 2019

[19] Ramdas Vankdothu,Dr.Mohd Abdul Hameed, Husnah Fatima" A Brain Tumor Identification and Classification Using Deep Learning based on CNN-LSTM Method" Computers and Electrical Engineering , 101 (2022) 107960

[20] Ramdas Vankdothu,.Mohd Abdul Hameed "Adaptive features selection and EDNN based brain image recognition on the internet of medical things", Computers and Electrical Engineering , 103 (2022) 108338.

[21] Ramdas Vankdothu,.Mohd Abdul Hameed,Ayesha Ameen,Raheem,Unnisa " Brain image identification and classification on Internet of Medical Things in healthcare system using support value based deep neural network" Computers and Electrical Engineering,102(2022) 108196.

[22] Ramdas Vankdothu,.Mohd Abdul Hameed" Brain tumor segmentation of MR images using SVM and fuzzy classifier in machine learning" Measurement: Sensors Journal,Volume 24, 2022, 100440

[23] Ramdas Vankdothu,.Mohd Abdul Hameed" Brain tumor MRI images identification and classification based on the recurrent convolutional neural network" Measurement: Sensors Journal,Volume 24, 2022, 100412 .

[24] Bhukya Madhu, M.Venu Gopala Chari, Ramdas Vankdothu,.Arun Kumar Silivery,Veerender Aerranagula " Intrusion detection models for IOT networks via deep learning approaches " Measurement: Sensors Journal,Volume 25, 2022, 10064

[25] Mohd Thousif Ahemad ,Mohd Abdul Hameed, Ramdas Vankdothu" COVID-19 detection and classification for machine learning methods using human genomic data" Measurement: Sensors Journal,Volume 24, 2022, 100537

[26] S. Rakesh [a], NagaratnaP. Hegde [b], M. VenuGopalachari [c], D. Jayaram [c], Bhukya Madhu [d], MohdAbdul Hameed [a], Ramdas Vankdothu [e], L.K. Suresh Kumar "Moving object detection using modified GMM based background subtraction" Measurement: Sensors ,Journal,Volume 30, 2023, 100898

[27] Ramdas Vankdothu,Dr.Mohd Abdul Hameed, Husnah Fatima "Efficient Detection of Brain Tumor Using Unsupervised Modified Deep Belief Network in Big Data" Journal of Adv Research in Dynamical & Control Systems, Vol. 12, 2020.

[28] Ramdas Vankdothu,Dr.Mohd Abdul Hameed, Husnah Fatima "Internet of Medical Things of Brain Image Recognition Algorithm and High Performance Computing by Convolutional Neural Network" International Journal of Advanced Science and Technology, Vol. 29, No. 6, (2020), pp. 2875 – 2881

[29] Ramdas Vankdothu,Dr.Mohd Abdul Hameed, Husnah Fatima "Convolutional Neural Network-Based Brain Image Recognition Algorithm And High-Performance Computing", Journal Of Critical Reviews,Vol 7, Issue 08, 2020