International Journal of Technology, Leadership and Sciences (IJTLS)

ISSN: 3107-9083
www.ijtls.com
Research Paper
DOI:https://doi.org/10.5281/zenodo.20799417

OPTIMIZED FEATURE SELECTION AND MSVM-BASED FRAMEWORK FOR AUTOMATED FAKE NEWS DETECTION

THOTA.VASAVI[1],DR.VUNNAVA DINESH BABU[2], Dr.CHAVA HARI BABU[3], R.VAMSI KRISHNA[4], D.SRIDHAR[5]

[1]M.Tech Student,RV Institute Of Technology,Chebrolu Mandal, Guntur District,Andhra Pradesh, India – 522212.

[2]Assistant Professor,RV Institute Of Technology,Chebrolu Mandal, Guntur District,Andhra Pradesh, India – 522212.

[3]Professor,RV Institute of Technology,Chebrolu Mandal, Guntur District,Andhra Pradesh, India – 522212.

[4]Assistant Professor,RV Institute Of Technology,Chebrolu Mandal, Guntur District,Andhra Pradesh, India – 522212.

[5]Assistant Professor,RV Institute Of Technology,Chebrolu Mandal, Guntur District,Andhra Pradesh, India – 522212.

ABSTRACT:

A significant issue in the contemporary interconnected world is the fast dissemination of disinformation and fraudulent material across many internet platforms. Disinformation may sway public perception, hinder democratic functions, and instigate pervasive uncertainty or anxiety. This paper presents an enhanced Multi-class Support Vector Machine (MSVM) classification method, using feature-based improvements to detect fraudulent news, therefore tackling the recognized difficulty. The system thoroughly comprehends incoming material by examining news articles for lexical patterns, syntactic structures, semantic representations, and contextual cues. To improve model efficacy and reduce computing complexity, these properties are refined by sophisticated selection methods. The MSVM classifier is taught to recognize not just binary outcomes but also several categories, including satire, partly true, and deceptive information, hence facilitating more nuanced and precise classifications. This hybrid method enhances detection precision, scalability, and flexibility across many data sources and languages. The system offers strong safeguards against the spread of disinformation via the use of machine learning and optimal feature engineering. It may be used in real-time contexts such as social media, fact-checking tools, and news monitoring systems. The suggested solution improves intelligent misinformation detection systems and addresses the shortcomings of current models.

Keywords: Fake News Detection, Multi-class Support Vector Machine (MSVM),Natural Language Processing (NLP), Text Classification, Misinformation Detection, Semantic Analysis, Lexical Features, Syntactic Features, Social Media Analysis, News Classification, Contextual Analysis.

Received Date: 5 June 2026**; Accepted Date:** 15 June 2026**; Published Date:** 20 June 2026;

I.INTRODUCTION

Digital news platforms and social media have revolutionized the dissemination and consumption of news in the modern era. Although current technology has made news and worldwide communication more accessible than ever, it has also encouraged the dissemination of disinformation. Fake news utilizes inaccurate or deceptive information to influence political choices, damage reputations, provoke social unrest, or alter public perception. The unchecked spread of false information endangers society, democracy,

public trust, and public health.Manual verification and rule-based filtering methods provide the basis of traditional strategies for detecting false news. Nonetheless, when faced with the continuous deluge of digital material, the majority of systems reveal inefficiency, need considerable effort, and display lethargy. Current algorithms have more difficulties in identifying modern fake news, since it often adopts intricate formats that closely resemble genuine news pieces.

This paper provides a feature-based optimized Multi-class Support Vector Machine (MSVM) classification system for the intelligent identification of counterfeit news, in light of these difficulties. The suggested method examines news articles for comprehensive information, including lexical, syntactic, semantic, and contextual components. Subsequently, we use sophisticated feature optimization methods to choose the most relevant properties, therefore improving classification accuracy and reducing computing complexity.The MSVM classifier employs optimum characteristics to classify news stories as authentic, satirical, partly true, misleading, or fake. The multi-class approach provides a more nuanced comprehension of categories of deceptive information than binary classifiers. The suggested system offers a scalable, accurate, and dependable solution for the real-time identification of disinformation across social media platforms, news websites, and fact-checking apps via the integration of machine learning and sophisticated feature engineering.

II.LITERATURE REVIEW

Kai Shu, Jiliang Tang, and Huan Liu conducted an extensive examination of data mining methodologies for detecting fraudulent news. The research investigates several feature extraction methodologies, including textual content analysis, user behavior analysis, and network-based features, with an emphasis on the attributes and difficulties associated with detecting fraudulent news. The researchers found that the accuracy of detecting false news markedly improved when content characteristics were combined with social contextual data.

William Yang Wang created the LIAR dataset to further the examination of lying, including approximately 12.8K carefully annotated political statements. The research assessed Logistic Regression and Support Vector Machines (SVM) among several machine learning methodologies, emphasizing precise categorization. Empirical research demonstrates that models with extensive features outperform in identifying false news.

Xinyi Zhou and Reza Zafarani elucidated the theoretical underpinnings and practical approaches for detecting fraudulent news in their comprehensive study. The researchers examined social and psychological ideas on misinformation and classified techniques for detecting false news into content-based and context-based categories. The results demonstrated that deep learning and machine learning are more essential in modern techniques for identifying bogus news.

Natali Ruchansky, Sungyong Seo, and Yan Liu devised the CSI (Capture, Score, Integrate) approach to detect disinformation. The technique integrates user behavior modeling using recurrent neural networks with the analysis of textual content. The CSI approach enhanced the accuracy and dependability of identifying false information by including temporal data and user interaction patterns.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi undertook a research examining varied linguistic and stylistic motifs in both deceptive and truthful news announcements. The study found emotion, subjectivity, hyperbole, and writing style as critical indications for identifying fraudulent news items.

Nicole J. Conroy, Victoria L. Rubin, and Yimin Chen analyzed methods for detecting fraud using linguistic and stylistic features. Their research assessed conventional machine learning classifiers, including Support Vector Machines and Naïve Bayes, using n-gram and syntactic data. The first stages of detecting false news demonstrated the efficacy of feature-based classification algorithms.

Shivani Jain, Nitin Kumar, and Deepak Gupta suggested the use of Recursive Feature Elimination (RFE) to improve features for machine learning-based false news identification. The revised feature selection method yielded enhanced classification

accuracy and decreased computing complexity in false news identification, hence improving the performance and efficiency of the Support Vector Machine classifier.

III.EXISTING SYSTEM

The primary methods for detecting disinformation include typical binary classification techniques using fundamental machine learning algorithms, including Logistic Regression, Naive Bayes, Decision Trees, and standard Support Vector Machines (SVM). These models mostly use content-based variables derived from news articles, including word frequency, TF-IDF scores, and sentiment polarity. Despite their simplicity and swift execution, these approaches often fall short when faced with the complex language structures and context-dependent meanings typical of false news items. This results from their reliance on superficial feature representations.

Furthermore, some modern systems neglect to use optimization methods for feature selection, resulting in high-dimensional input data that may include noise or extraneous information. This impairs the models' capacity to generalize to new data, often leading to overfitting or underfitting. Furthermore, these algorithms often utilize conventional support vector machines (SVMs) for binary classification, overlooking multi-class categorization and the complexities of news propagation dynamics, source credibility, and writing style—factors crucial for effectively differentiating false news from genuine news.

The lack of integration with comprehensive data sources, such as social media platforms or real-time data, is a major problem with existing systems. The majority of systems rely on static datasets and are incapable of adjusting their algorithms to new forms of deceit. The lack of strong interpretability and explanations provided by most systems diminishes their dependability for public use. Therefore, while current methods for detecting fake news have laid a groundwork, improved models are essential to increase accuracy and scalability in practical applications through advanced classification techniques such as Multi-class SVMs and optimized feature selection.

DISADVANTAGES

- Unfortunately, most modern algorithms neglect complex semantic and contextual nuances in news articles, since they emphasize fundamental textual metrics like word count, TF-IDF, and sentiment polarity.
- Unmitigated high-dimensional noisy data, without dimensionality reduction and feature selection, results in reduced performance, prolonged processing time, and heightened chances of overfitting or underfitting.

IV.PROPOSED SYSTEM

The suggested method seeks to overcome the shortcomings of current fake news detection systems by using a feature-rich, optimized Multi-class Support Vector Machine (MSVM) classification model. This system utilizes sophisticated Natural Language Processing (NLP) methods to extract both profound and superficial attributes from news articles. A thorough comprehension of the article's content may be achieved by the analysis of lexical features (TF-IDF, n-grams), grammatical structures, readability metrics, and sentiment attributes. The model can detect nuanced indicators of disinformation, including sensational language, inconsistent writing, and deceptive strategies, owing to its extensive feature set.

Principal Component Analysis (PCA), Recursive Feature Elimination (RFE), and Genetic Algorithms (GA) are approaches for feature optimization aimed at enhancing model performance and minimizing noise. These solutions improve classifier efficacy while minimizing computing expenses by identifying the most relevant and meaningful characteristics. Thereafter, a Multi-class Support Vector Machine is trained using the optimized feature set rather than a standard binary SVM. The MSVM may categorize news into many classifications, such as Real, Fake, Satire, and Misleading, to augment its adaptability for a range of practical applications. To prevent overfitting, it is crucial to adjust the model's hyperparameters using a framework like GridSearchCV.

The finalized solution will exhibit scalability and interoperability with existing systems that aggregate news and provide real-time social media monitoring for users. Accuracy,

Precision, Recall, F1-Score, and AUC-ROC are common measures used to assess dependability and robustness. To improve the clarity of categorization judgments for users, the incorporation of explainability tools such as SHAP or LIME may be included in the system. The suggested approach guarantees improved accuracy, broader scope, and practical use, exceeding existing false news detection techniques by integrating optimized characteristics with a robust classification engine.

ADVANTAGES

- In integrating statistical and semantic text analysis, the system improves its capacity to identify subtle language patterns that would be undetectable by human examination.
- The framework's scalability and agility allow the effortless incorporation of new datasets, features, and classification models for future improvements.

V.SYSTEM MODEL

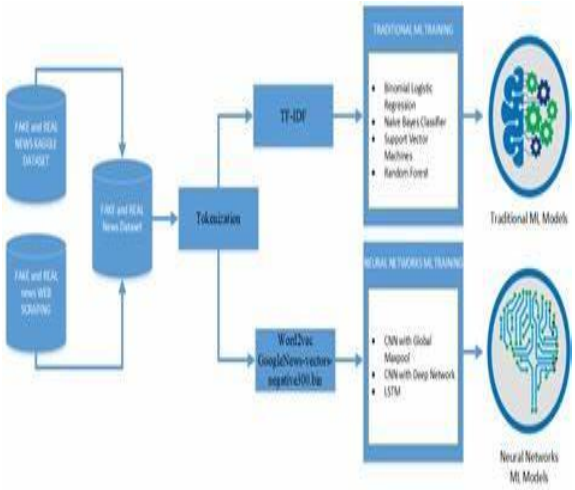


Fig.1 System Model

Figure 1 above depicts the process flow of a system using deep learning and conventional machine learning to categorize false information. The first step in the processing and analysis of false and authentic news databases is collecting them from several sources into a unified dataset. To improve the efficacy of text analysis, information is subjected to preprocessing and tokenization, which involves breaking down news items into smaller units such as words or phrases.

Following the preprocessing phase, two distinct feature extraction techniques are used. Traditional machine learning techniques use TF-IDF (Term Frequency-Inverse Document Frequency) to quantitatively assess the significance of words in a dataset by converting textual data into feature vectors. A range of classifiers, including as Random Forest, Support Vector Machines (SVM), Logistic Regression, and Naïve Bayes, are trained using these feature vectors.Concurrently, deep learning models receive word vectors obtained from sources like Google News using Word2Vec embeddings. Neural network topologies, like CNN, CNN with Global Max Pooling, and LSTM networks, use these embeddings as input, which encode the contextual and semantic links between words.

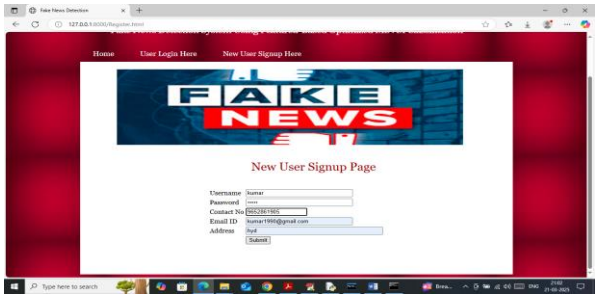
Ultimately, models trained on processed data may effectively detect bogus news items, irrespective of their reliance on classic machine learning or neural network techniques. The systematic use of statistical feature extraction with semantic deep learning representations improves the efficacy and precision of false news detection.

VI.MODULES

To implement this project we have designed following modules

1) New User Sign up: using this module user can sign up with the application
2) User Login: using this module user can login to system
3) Load Fake News: using this module user can upload fake news dataset to application and then apply NLP algorithms on loaded news text data to remove stop words and special symbols
4) Run MPCA & Firefly Features Selection: cleaned news data will be input to MPCA algorithm to extract features and then apply Firefly algorithm to select relevant features. Selected features will be split into train and test where application using 80% data for training and 20% for testing
5) Run MSVM Algorithm: 80% training data will be input to MSVM and LSTM algorithm to train a model and this model will be applied on 20% test data to calculate prediction accuracy
6) Predict News: using this module either user can enter NEWS data or upload news file and then system will apply all algorithms to classify that news or False or True.

VII.SCREENSHOTS



The user is entering sign-up information on the aforementioned screen and then hits a button to go to the next page.

After completing the user registration, choose the 'User Login' option to proceed to the next page.

Upon signing in, the user will be sent to the following screen seen above.

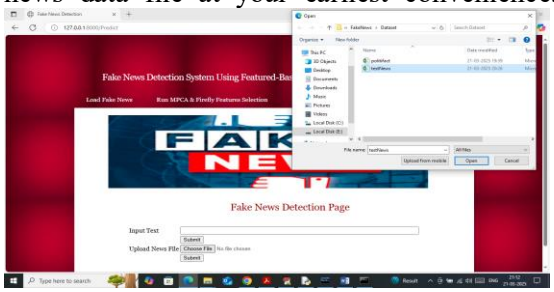
The dataset shown above has a first column with news data and a second column with class labels, denoting either FALSE or TRUE. Select the 'Run MPCA and Firefly' option to extract and select features from the dataset, which will direct you to the subsequent page.

The first two lines of the screen indicate the quantity of records and characteristics included in the dataset. Subsequently, the MPCA extracted 300 features, whereas Firefly identified 160 key features. Subsequently, choose the 'Run MSVM Algorithm' link to begin the training of the algorithms, directing you to the subsequent page

The screen above presents a table including the accuracy, precision, recall, and F1 score of the proposed MSVM versus the existing LSTM approach. The table demonstrates that MSVM acquired an accuracy of 98%, whereas the current LSTM achieved an accuracy of 87%. In the confusion matrix, the x-axis represents Predicted Labels, while the y-axis indicates True Labels. The yellow and light green boxes along the diagonal represent the number of accurate forecasts, whilst the blue boxes indicate incorrect guesses. The x-axis of the bar graph represents method names, whilst the y-axis signifies accuracy and other metrics, shown by uniquely colored bars. Select the 'Predict News' hyperlink to get to the next page.

In the previously described screen, I inserted news data into the text field and then pushed the button to go to the next page.

The aforementioned news data has been categorized as 'Fake,' and you may likewise assess any other news article. Kindly deliver the news data file at your earliest convenience.



Upload the 'testNews.csv' file at the top of the screen, then select 'Open' and subsequently the second submit button to go to the next page.

The first column exhibits news text data, whilst the subsequent column reveals classification results as 'False' or 'True.'

VIII.CONCLUSION

The False News Detection System provides a comprehensive and efficient remedy to the pervasive issue of misinformation on digital platforms. This is achieved by a feature-optimized Multi-class Support Vector Machine (MSVM) classification. This approach enhances classification accuracy and reduces computational complexity by using many textual attributes, including syntactic, semantic, and lexical dimensions, with advanced feature optimization methods. Employing MSVM as the primary classifier enables the model to transcend basic binary classification, offering subtle insights by skillfully distinguishing among several categories, including authentic news, misinformation, satire, and fraudulent content. This enables the system to be very adaptable and relevant to many real-world scenarios where diverse forms of deception may exist. The model's robustness and capacity to generalize across diverse datasets are ensured by stringent training and validation techniques. These tactics ensure that the model can adjust to changing news styles and sources.

Implementing the system in social media monitoring, news aggregation, and fact-checking platforms is advantageous due to its scalable architecture and ability for real-time integration. These forums are crucial for the rapid identification of deception. Utilizing explainability tools enhances value by elucidating decision-making, therefore augmenting user trust and fostering acceptance among journalists, academics, and the general populace. This proposed technique represents a significant advancement in combating misinformation, since it addresses fundamental issues present in existing systems, including elevated processing costs, limited feature sets, and binary classification constraints. Ultimately, it significantly aids in combating misinformation, disseminating correct information, and fostering informed discussion in an increasingly digital and interconnected landscape.

IX.FUTURE ENHANCEMENTS

The current Fake News Detection System, using feature-based optimized MSVM classification, offers a solid foundation for rapid and precise detection; nevertheless, further enhancements might improve its effectiveness. An effective strategy involves using deep learning approaches such as GPT, BERT, or RoBERTa, which are grounded on transformer architecture. These models excel at identifying nuanced disinformation and intricate falsehoods that traditional classifiers may overlook, because to their superior understanding of textual context and semantic relationships.

Utilizing several methodologies to assess news articles is a possible avenue for future research. Deceptive language is just one facet of misinformation; manipulated images, videos,

and audio recordings are also prevalent. Incorporating the ability to process and analyze multimedia material with textual information may improve the system's recognition and understanding skills. The system would detect inconsistencies across diverse material formats and highlight them to enhance overall reliability via the use of computer vision and natural language processing techniques.

Developing frameworks capable of continuously monitoring social media feeds or news websites in real time for deception represents a viable strategy to enhance scalability and deployment. Lightweight models that enhance accuracy and reduce latency, together with an efficient data pipeline design, are essential for this objective. To effectively handle the substantial data volume, it may be essential to explore cloud-based solutions or edge computing methods for the implementation of automated fact-checking and prompt alerts without compromising system performance.

Additionally, a vital aim for the future is to enhance the clarity and comprehensibility of categorization determinations. Understanding the rationale for designating a news report as fake is beneficial for users, including both journalists and casual readers. Integrating advanced explainable AI (XAI) techniques that provide transparent elucidations and delineate the contributions of certain textual elements may be crucial for future advancement. This will assist fact-checkers in verifying and disproving inaccuracies, while simultaneously enhancing user trust.

The last phase involves using continuous learning and domain adaptation methodologies to enhance the system's resilience and adaptability. Given the rapid evolution of misinformation's structure, content, and transmission methods, the detection system must continuously adapt by incorporating new data while preserving its existing knowledge base. To achieve enhanced detection accuracy in dynamic environments and diverse cultural and linguistic settings, approaches such as adversarial training, incremental learning, and transfer learning must be considered.

X.REFERENCES

1. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22–36. https://doi.org/10.1145/3137597.3137600
2. Wang, W. Y. (2017). “Liar, liar pants on fire”: A new benchmark dataset for fake news detection. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (pp. 422–426). https://doi.org/10.18653/v1/P17-2067
3. Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys (CSUR), 53(5), 1–40. https://doi.org/10.1145/3395046
4. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018
5. Kaur, H., & Singh, A. (2020). Fake news detection using machine learning approaches: A review. International Journal of Computer Applications, 176(23), 1–7. https://doi.org/10.5120/ijca2020920249
6. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?” Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778
7. Jain, S., Kumar, N., & Gupta, D. (2021). Fake news detection using recursive feature elimination and support vector machine. International Journal of Advanced Computer Science and Applications, 12(5), 401–408. https://doi.org/10.14569/IJACSA.2021.0120548
8. Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, 52(1), 1–4.

https://doi.org/10.1002/pra2.2015.145052010082

9. Ruchansky, N., Seo, S., & Liu, Y. (2017). CSI: A hybrid deep model for fake news detection. In Proceedings of the 2017 ACM Conference on Information and Knowledge Management (pp. 797–806). https://doi.org/10.1145/3132847.3132877
10. Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 2931–2937). https://doi.org/10.18653/v1/D17-1317.