International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 12 | Dec 2025 www.irjet.net p-ISSN: 2395-0072

A REVIEW OF FAKE NEWS DETECTION USING NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING

Khushboo Yadav[1], Mr. Rajesh Sharma[2]

[1]*Master of Technology, Computer Science and Engineering, Sagar Institute of Technology and Management, Barabanki, India*
[2]*Assistant Professor, Department of Computer Science and Engineering, Sagar Institute of Technology and Management, Barabanki, India*

Abstract - *The signal increase in digital content and the overwhelming domination by social media platforms have contributed to the massive spread of fake news, and subsequent exposure to those poses significant risks to societies, political regimes, and confidence. In the proposed review paper, the trajectory of the development of fake news detection algorithms is thoroughly discussed, and machine learning-based algorithms and methods based on Natural Language Processing (NLP) are singled out as the methodology of their paramount study. The evolution of the rule-based systems to supervised and unsupervised models and models using keyword detection, Naive Bayes, Support Vector Machines, Logistic Regression, Random Forest, and clustering algorithms has been traced as a backdrop of the study. It is stressed that the crucial place belongs to the feature engineering procedure, like TF-IDF, Bag-of-Words, and n-grams, as it increases the model accuracy.*

Advanced Deep Learning (DL) approaches such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) and Transformer based models and especially the BERT approach are further discussed in the paper with its overall ability to capture semantic and contextual understanding in the text being highly superior to the other techniques. Adoption of NLP techniques, such as tokenization, part-of-speech tagging, lemmatization, sentiment analysis, stance detection, and contextual word embedding's such as Word2Vec and Glo Ve, is critically reviewed concerning their contribution to the performance and feature extraction. NLP-ML/DL hybrid models and ensemble learning are examined to analyze how they can be used to fulfill the requirements of complex detection tasks.

Key Words: Fake News Detection, Natural Language Processing (NLP), Machine Learning (ML), Deep Learning (DL), Feature Engineering, Misinformation Analysis, Hybrid Models.

1. INTRODUCTION

1.1 Fake News and Its Impact

Fake news can be seen as any misleading or false information propagated as valid news. As opposed to the misinformation, the fake news is usually shared to deceive, shape opinions, and harness political, financial, or social benefits. Fake news is a significant issue affecting the world, particularly since the introduction of digital media. The effects of it are also multidimensional, as it might affect political elections, bear violence, cause social volatility, ruin the reputation, and decrease the overall trust in the valid news sources. Fake news, mainly when it's spread during emergencies, has highlighted the necessity of effective detection dynamics.

1.2 Role of Social Media in Dissemination

Introducing social media like Facebook, Twitter, Instagram, and What's App has transformed how information is exchanged and taken in. Although such platforms allow instant communication and access to democratized information, they have become the global hotbeds of itinerant fake news. The social media practices of algorithm-based content delivery are more likely to promote engagement, which will often enlarge on sensational or emotionally passionate articles or commentaries that may or may not be truthful. Compared with traditional media, where checks are in place by editors and other standards regulating media, social media has no elaborate fact-check procedures and, therefore, news that is not genuine can easily be published to a potential mass audience in no time. The viral nature of the content published on these online platforms offers great challenges to controlling misinformation. It evidences the necessity of technological solutions that help monitor and analyze social media information in real time.

1.3 Challenges in Detecting Fake News

The detection of fake news has some inherent challenges. To begin with, the fake news stories resemble regular news articles with realistic headings, authentic sources, and real distortion of facts that even expert deputies can hardly tell apart from the genuine information. Second, the linguistic patterns of fake news are diverse and, in many cases, context-sensitive, and, therefore, traditional methods based on keywords or rules are not extensive enough. Moreover, the current misinformation activity, including bot activities, deep fakes, and mechanisms of misinformation campaigns, develops quickly and challenges the detection process even more. The multilingualism of content posted in social media

and the ever-new narratives and slang introduced necessitate active and flexible detection models. Besides, the absence of massive, large-scale, labeled data is a serious problem in training successful machine learning models.

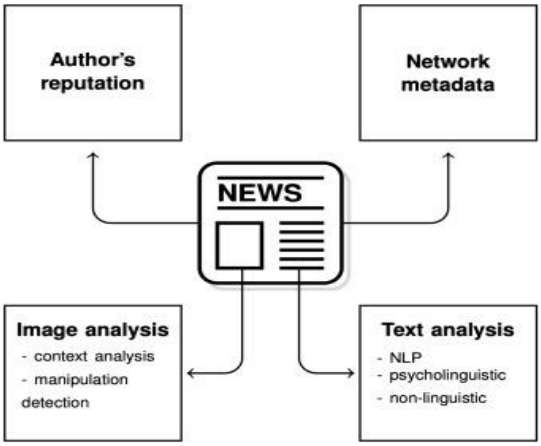

Figure 1: Analysis of the NEWS Components.

1.4 Importance of Automated Detection Methods Using NLP and Machine Learning

Manual detection methods cannot possibly be sufficient and effective due to the scale and pace of spreading fake news. The promising methods of identifying this urgent issue are relational finding systems based on Natural Language Processing (NLP) and Machine Learning (ML) technologies. NLP allows the retrieval and processing of the textual characteristics, namely syntax, semantics, sentiment, and writing style, which play an essential role in differentiating fake news and genuine messages. Classification of news articles based on the learnt patterns may be attained by training a machine learning algorithm, especially the supervised learning models, on annotated datasets. Moreover, recent breakthroughs in deep learning technology, transformer-based models (like BERT and GPT) have made modern detection systems much more potent as regards comprehending and processing human language, optimizing the possibility of detecting fake news, which makes the automated phony news detection more accurate and much more scalable. NLP/ML integration, therefore, has the potential to develop formidable and real-time detection systems that can curb the effects of fake news in society.

1.5 Objectives of the Review

The main aim of such a review is to review and synthesize the current studies regarding fake news detection methods that rely on using Natural Language Processing and Machine Learning. The research paper seeks to discuss and establish an in-depth review of the conventional machine learning algorithm, state-of-the-art deep neural networks, and hybrid models, which combine different algorithms to enhance classification accuracy. The review will also endeavor to highlight the application of NLP in feature extraction and language comprehension in the context of fake news classification. This review aims to define the existing research gaps by analyzing strengths, weaknesses, and areas of practical application of various detection methods, thus providing some insights into the further development of the detection system of fake news. Finally, the paper can be of great value to scholars and practitioners in combating the fact that misinformation has also become difficult due to the digital age.

2. LITERATURE REVIEW

2.1 Early Approaches to Fake News Detection

2.1.1 Rule-Based Systems

During the early days, when fighting misinformation and fake news was at its very beginning, one of the first technical solutions was to use rule-based systems. They are based on linguistic or syntactic rules created manually to determine whether the contents of the news articles are dubious or deceitful. These rules usually rest upon observable characteristics, including exaggerated language, sensational headlines, and some noticeable keywords or anomalies in style, which do not belong to the usual standards of journalists. Rule-based systems run on a collection of pre-defined heuristics, which are commonly developed by the local experts in a field or linguistics to identify patterns of the possible occurrence of a lie being present.

Credibility indicators might be considered one of the excellent examples of rule-based detection, therefore, flagging a possible misinformation based on using capital letters too often, or using exclamation marks, hyperbolized expressions, and emotively-charged words (Ruskin et al., 2017). Earlier fake news detection systems have tried to manually capture similarities between the fake news articles and identify heuristic patterns linguistically. Nevertheless, although simple to interpret and implement, the rule-based systems experience several fatal flaws. They are too rigid and unable to adjust to the new language of the fake news or sense the context in various fields. Also, these systems entail constant manual requesting, and thus, it is more difficult to scale these systems in a dynamic environment like online news (Conroy, Rubin, & Chen, 2015).

2.1.2 Keyword Detection and Manual Flagging

The other early technique used in fake news identification was keyword detection and manual flagging, where certain words or phrases used primarily in misinformation were used as identifiers of bogus stories. This strategy frequently included the development of a dictionary of illicit keywords/sources that tend to publish. For example, phrases such as shocking, miracle, or you will not believe were frequently included in the targeting as they refer to click bait or misleading titles (Potthast et al., 2017).

In turn, manual flagging used humans as moderators or fact-checkers who controlled the content and marked it as fake or authentic. Fake news filtering on several social media platforms, particularly during the early stages of fake news management, was left to community reports and fact-checking teams that were manually driven. Nonetheless, keyword-based detection and manual flagging are dynamic processes in their effectiveness, even though they are limited. Keywords, at times, do not carry contextual knowledge and could bring up a situation of false positives or negatives. Similarly, flagging by hand is bandwidth-demanding, time-consuming, and impractical in processing the vast data volume posted in digital platforms daily (Zhou & Zafarani, 2018).

However, as much as they constituted the backbone, these historical solutions emphasized the difficulty of identifying fake news, and the shift toward a more flexible and autonomous solution was necessitated, leading to the concept of incorporating Machine Learning and Natural Language Processing in future research.

2.2 MACHINE LEARNING-BASED APPROACHES

Weak performance of rule-based and keyword-driven methods of fake news detection gave reason for researchers to seek more dynamic and data-driven approaches. The introduction of Machine Learning (ML) techniques has had a profound impact in that the models can be trained on data patterns, allowing them to make predictions on instances they have never seen without having to be coded explicitly. Fake news detection using ML models can broadly be grouped into supervised and unsupervised learning tasks, with pros and cons in analyzing the textual data.

2.2.1 Supervised Learning Models

Supervised learning entails algorithm training using labeled datasets, whereby the entries of news content are labeled as fake or real. The model's training involves learning the relationship between textual features and their corresponding label. It enables the model to classify new, unseen articles based on learned patterns.

The Naive Bayes classifier is one of the first and most common supervised models of detecting fake news. Based on the probabilities of learning, Naive Bayes performs the operations based on the independence of features. As a result, it has an advantageous computational cost and has found favorable applications in the text classification problem. However, even by being constructed, Naive Bayes has demonstrated competitive rates on fake news detection in situations where effective feature extraction methods were used (Perez-Rosas et al., 2018).

The other standard algorithm is the Support Vector Machine (SVM) algorithm, which uses the optimal hyper plane to partition data points belonging to various classes in a high-dimensional space. One of the most promising binary classification settings where SVMs have the highest chance of success is their use in fake news identification, since they utilize the more sparse representations of data, such as TF-IDF vectors (Shu et al., 2017).

Logistic regression is a linear model where the probability of a binary outcome is estimated as a function of supplied features. Although the model is very straightforward, yields interpretable outcomes, and is scaled to high-dimensional text, it is suitable in association with relevant feature engineering approaches (Ahmed et al., 2017).

The ensemble approach, such as the Random Forest classifier, has also been heavily used. Random Forests overcome this tendency because they use several decision trees constructed during training, allowing them to decrease over fitting and improve predictive Precision. After all, the predictions of several trees are averaged. Their resistance against noisy data and feature significance estimation functionalities might offer them the desired results when used in a fake news classification exercise (Kaliyar et al., 2021).

2.2.1.1 Feature Engineering Techniques

Feature engineering plays a central role in the achievement of machine learning models when used in text-based fake news detection. The conventional models are built based on transforming raw text into numerical representations that algorithms may use. The popular methods are:

Bag-of-Words (Bow): Represents a text based on the number of times words are used, without caring about grammar and the order of words. Nevertheless, Bow is good at picking superficial textual patterns, even though the technique is simple.

Term Frequency-Inverse Document Frequency (TF-IDF): A statistical ratio, which indicates the essentiality of a word to a document as contrasted to a corpus. TF-IDF decreases the influence of words used habitually and increases that of the specific ones that provide more discrimination of the features.

N-grams: They take a sequence of n words (e.g., bigrams, trigrams) and can lead the model learning to consider a word in context and word co-occurrence, which helps to detect movements referring to specific phrases common in fake news (Rashkin et al., 2017).

Table 1: Machine Learning Models.

Model	Key Feature	Strength	Limitation
Naive Bayes	Probabilistic, assumes independence	Fast, efficient with text data	Assumes feature independence
SVM	Hyper plane-based separator	Effective with high-dimensional data	Computationally expensive with large datasets
Logistic Regression	Linear classification	Interpretable and scalable	Limited to linear relationships
Random Forest	Ensemble of decision trees	Robust to over fitting, provides feature importance	Maybe less interpretable

2.2.2 Unsupervised Learning Models

In contrast to its supervised counterparts, the unsupervised learning models do not require labeled data. Such algorithms interrogate the internal structure of the information to seek a pattern, grouping, or anomalies. Unsupervised approaches are less typical in fake news detection than supervised learning; yet one analysis built on them, complex neoclassical centiloid clustering of news stories, and latent topic discovery of data sets.

Among the methods is clustering techniques like the K-means clustering, which classifies the data into various groups depending on the similarity of the features. In fake news, clustering allows for the division of news articles under a common language or style, which could separate genuine news sources from counterfeit sources (Zhang & Ghorbani, 2020). Nevertheless, clustering is insufficient to mark the clusters, and thus it is restricted to specific applications and classification-type tasks.

Topic modeling, e.g., Latent Dirichlet Allocation (LDA), is another primary method in the unsupervised toolbox. Topic modeling is the solution to the structural analysis of the set of writings, to track down the latent thematic organization. LDA models presume that documents can be considered as a combination of topics, and in turn, topics are a mixture of words. Examining these distributions, LDA can identify popular topics in fake news articles, whether sensationalism or anything related to a conspiracy, helping analysts know about the misinformation narrative (Bondielli & Marcelloni, 2019).

2.3 DEEP LEARNING-BASED APPROACHES

As is the trend with more powerful computers and the advent of massive data, deep learning methodologies have transformed how natural language is processed and, by extension, fake news detection. In contrast to the classical machine learning models, where manual extraction of features is indispensable, deep learning models automatically extract hierarchical representations in raw data. Such ability has considerably enhanced the performance and generalization of fake news detection systems.

2.3.1 Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNNs) have been traditionally used in image processing challenges, but have been more useful/practical in text classification, such as the case of fake news detection. The action of CNNs is convolutional filtering of the data presented to take up local information in the form of n-grams or phrase patterns. CNNs are desired because they use local dependencies and spatial hierarchy of data that appear in text data, such as stylistic phrases and deceptive language structure, which tend to make fake news articles. Several studies have already shown that with large datasets of phony news, CNN-based models can perform better than the conventional ones by automatically learning the necessary features through data (Wang, 2017). Nevertheless, CNNs can have problems capturing long-term relationships in sequences, which limits their sensitivity when more contextual information is needed.

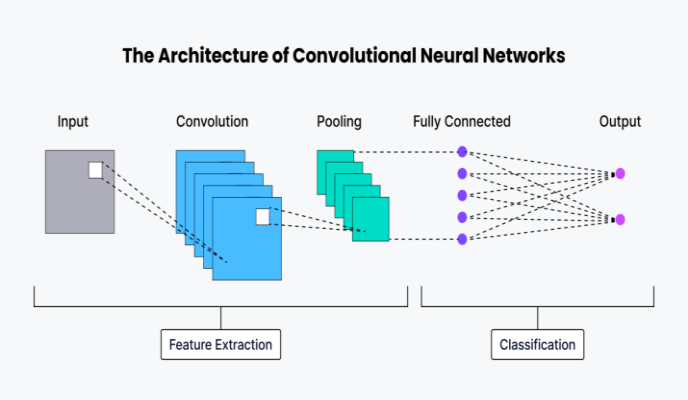

Figure 2: Convolutional Neural Networks (CNN)

2.3.2 Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM)

The Recurrent Neural Networks (RNNs) were devised to overcome the drawback of incorporating the sequential dependencies in text. The RNNs are made to analyze sequence data where the information of a specific time step is transferred into the next time step using the hidden state within the network. RNNs have helped in fake news detection as RNNs allow the model to take into account the sequencing of words and contextual relations, which is instrumental in understanding a news article's semantics.

But the traditional RNNs are vulnerable to vanishing and exploding gradient issues and thus cannot learn long-range dependencies well. The problem was alleviated when the Long Short-Term Memory (LSTM) networks were developed (Hoch Reiter & Schmidhuber, 1997). LSTM networks add gating functions, such as input, output, and forget gates, to control what information passes through a long sequence and enable the model to keep or drop information as needed.

Research has demonstrated that semantics and syntactic patterns applicable in detecting fake news can be captured using the LSTM-based models, such as recurring narratives or sentiment cues. Ruch ansky et al. (2017) also suggested a hybrid model, such as CSI (Capture, Score, Integrate), which included LSTM with features of user engagement and proved that it was feasible to improve fake news detection by models that used both content distribution patterns and propagation patterns sequentially.

2.3.3 Transformers and BERT-Based Models

The latest NLP research achieved the most significant breakthrough when the Transformer architectures were introduced (Vaswani et al., 2017). In contrast to RNNs, Transformers can consider all tokens in an arrangement while appointing them to model global dependencies beyond the existence of persisting relationships. The architecture significantly increased training efficiency and the models' performance in multiple NLP tasks.

Based on Transformer, the Bidirectional Encoder Representations from Transformer (BERT) became one of the robust pre-training language models with an ability to learn contextual associations of the textual data in either direction (Devlin et al., 2019). The advantage of BERT is its capability to view left and right context identifiers of the sentence being trained, which helps it gain subtle meanings, thus identifying the slightly tricky hints on false news pieces.

BERT-based models have reached state-of-the-art performance regarding fake news detection, where pre-trained knowledge on large corpora and fine-tuning on given fake news datasets is applied. Fake BERT, developed by Kaliyar et al. (2021), is a model updated by fake news datasets, and compared to the traditional machine learning models and the earlier deep learning method, it was much more effective. The ability of BERT to engage in contextual understanding and its versatility to engage in domain-specific fine-tuning has created a new bar in passively detecting fake news.

Table 2: Comparative Overview of Deep Learning Models for Fake News Detection.

Model	Key Mechanism	Strengths	Limitations
CNN	Convolution over word embedding's	Captures local features effectively	Limited in modeling long-range dependencies
LSTM	Sequential processing with memory gates	Models sequential and contextual dependencies	Computationally intensive for long sequences
Transformer/BERT	Self-attention mechanism	Captures global dependencies and context bidirectionally	Requires ample computational resources for training

2.4 NLP TECHNIQUES IN FAKE NEWS DETECTION

Natural Language Processing (NLP) is the foundation stone for automated fake news detection systems since it helps the machine process, analyze, and comprehend human language. Even though implementing the Machine Learning and Deep Learning models to the problem would create the computational environment, the success of the models would greatly rely on the language processing methods used in the preprocessing and obtaining features. NLP methods are also employed on raw text data so that meaningful linguistic, syntactic, and semantic features can be gathered to enhance the sensitivity of the automated systems.

2.4.1 Tokenization, Part-of-Speech (POS) Tagging, and Lemmatization

Tokenizing can be identified as one of the most basic procedures in NLP-aided textual analysis, as it includes dividing a specific text into smaller segments, including words, phrases, or sentences. The information analysis of text using models could be implemented on tokenization, which allows for treating each text element independently and identifying such properties as the frequency of words or n-grams. In detecting fake news, tokenization is necessary to transform the unstructured text into a structured form, which could be further submitted to algorithms (Bird, Klein, & Lopper, 2009).

Special software, Part-of-Speech (POS) tagging, adds grammatical labels (noun, verb, adjective) to each token due to the tokenization. The POS tagging can be used in syntactic parsing and stylistic or grammatical patterns likely to signal fake news content can be discovered. For example, the proportion of adjectives or sensational verbs that prompt emotional reactions of the audience presented in counterfeit news articles may be higher than in other articles (Rashkin et al., 2017).

The other important NLP tool is Lemmatization, which involves reducing words to their base or dictionary form. It is called the lemma. Compared to stemming, wherein words tend to be shortened loosely, Lemmatization involves adjustments made to the linguistic meaning of the word that has been changed. The procedure assists in normalizing textual data, lowering dimensionality, and enhancing the consistency of features obtained by a text corpus. Lemmatization is beneficial in identifying the similarities of the words that share the same word semantic meaning and increase feature homogeneity in the case of fake news detection.

2.4.2 Sentiment Analysis and Stance Detection

Sentiment analysis is one of the more mainstream NLP tasks in which emotion or opinion expressed in text is detected and classified. It is especially applicable in fake news detection since fake news materials frequently use emotionally charged language to persuade readers regarding perception manipulation or encouraging extreme responses. By modeling such sentiment polarity, that is, categorizing a statement as positive, negative, or neutral, detection systems can be built to determine whether the provided emotional tone of the material is different from the accepted news reporting conventions (Giachanou, Rosso, & Crestani, 2019).

Stance detection takes this one step further by identifying the positioning or attitude of a text concerning a given claim, subject, or entity. Compared to sentiment analysis, which focuses on expressing an emotional temper, stance detection determines whether the material supports, opposes, challenges, or remains neutral regarding a given statement. It is beneficial in fake news detection, where one examines articles or social media posts that cite known misinformation. Identification of the stance will show the patterns in which fake news tends to favor certain conspiratorial or misleading stances most of the time (Hanselowski et al., 2018).

Sentiment analysis and stance detection help contribute to a deeper semantic analysis of text, and models capable of identifying fake news can pick up nuance that would not be apparent at the surface.

2.4.3 Contextual Word Embedding's (e.g., Word2Vec, Glove)

The conversion of textual data into numerical format must be a step needed by computational models, and word embedding's having seen an advantageous technique in this regard. Word embedding's contrast to more classical feature extraction, such as Bag-of-Words or TF-IDF, in that they operate on the continuous vector space framing, rather than the discrete word-counting space.

One of the most widely spread word embedding approaches is Word2Vec, discovered by Mikolov et al. (2013). It resorts to shallow neural networks to store word associations given a context window. It will generate a dense vector in which semantically related words are separated by a smaller amount in the embedding space. Word2Vec allows for a better classification of the different classes of fake news by aiding the model in identifying contextual use of words and nuances of the semantic connection with other words. Therefore, it increases the accuracy of the classification model.

Glove (Global Vectors for Word Representation) was designed by Pennington et al. (2014), who tried to integrate Matrix factorization approaches with the local context learning into a method of word vector generation. In contrast to the position of word2vec, which works on local context, Glove uses statistical information on a global scale about the whole corpus, making it able to capture indirect and direct mapping of words to each other. Glove embedding's have been successfully applied in different fake news detection papers to increase the semantic understanding of textual data.

Table 3: NLP Techniques in Fake News Detection.

NLP Technique	Purpose	Application in Fake News Detection
Tokenization	Splits text into meaningful units (tokens)	Prepares text for further analysis
POS Tagging	Assigns grammatical categories to tokens	Identifies syntactic patterns and stylistic cues
Lemmatization	Reduces words to their base form	Normalizes text and reduces feature space
Sentiment Analysis	Analyzes the emotional tone of text	Detects emotionally charged or manipulative language
Stance Detection	Determines the position of text on a claim	Identifies bias and alignment with misinformation
Word2Vec	Learns word associations from context	Captures semantic similarities for classification
GloVe	Learns word embeddings from global statistics	Enhances semantic understanding in detection models

2.5 HYBRID APPROACHES

The challenges associated with detecting fake news and the rich and changing malign elements have stressed the need for more than a single machine learning or Deep learning approach. To answer this, researchers have come to question

hybrid techniques, a form of computational method that integrates several different techniques and therefore uses the strong points of each. The Natural Language Processing (NLP) based methods tend to be based on Machine Learning (ML) or Deep Learning (DL) and usually implement the ensemble learning approach so that they reveal high accuracy, are scalable, and robust.

2.5.1 Combining Machine Learning/Deep Learning with NLP Techniques

Among the most powerful tactics of hybrid strategies is combining the smooth integration of feature extraction based on NLP with higher ML or DL classifiers. It starts with preprocessing the textual data through traditional NLP-based techniques such as tokenization, part-of-speech tagging, lemmatization, sentiment analysis, and stance detection. Such methods help extract functional linguistic and semantic patterns from unformatted text, which are instrumental in determining the difference between real and fake content.

After the extracted features are passed through machine learning algorithms such as Support Vector Machines (SVM), Random Forests, or Logistic Regression, they are used to perform classification activities, researchers like Ahmed and colleagues (2017) have displayed that a relatively good performance increase of fake news detection systems is achieved when NLP-based features (e.g., TF-IDF, sentiment scores) are paired with supervised learning methods.

Similarly, another practical method, on par with pre-trained word embedding's, is that of deep learning-based models, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs), which have been combined effectively to add deeper semantic comprehension. For example, Ruchansky et al. (2017) suggested the content-based features, user engagement patterns, and temporal information through a combination of neural networks and feature extraction techniques via CSI model, a hybrid deep learning framework. This strategy performed better than others that use content features, and one can note that such an approach promises efficiency in multidimensional factors of fake news detection.

Furthermore, the hybrid systems have also been expanded to analyze metadata, including the source's credibility, time of publication, and areas in which the user uses it in an orderly manner. Such systems use any news article's textual and contextual background, thus giving a more comprehensive analysis.

2.5.2 Ensemble Learning Models

Ensemble learning is the other hybrid technique that integrates two or more classifiers to create a more accurate and robust prediction model. The concept of ensemble methods presupposes that combining the output of multiple models will allow the system to minimize bias, variance, and possibilities of over fitting, eventually enhancing the generalization performance.

Fake news detection has several ensemble methods:

Bagging (Bootstrap Aggregating): In this approach, several replicates of the same model are trained on individual random samples of the training data, and their predictions are averaged, e.g., by majority voting. However, Random Forest is one of the standard examples of a bagging ensemble used to achieve relevant results in fake news detection (Kaliyar et al., 2021).

Boosting: Boosting also differs from bagging in that, whereas bagging trains models in parallel, boosting sequentially trains models with the aim of the next model in the sequence correcting any mishaps made by the previous model. Such algorithms as Ada Boost and Gradient Boosting, among others, have been applied to achieve higher predictive performance of fake news classifiers via paying attention to cases that are not easily classified (Shu et al., 2017).

Stacking: In this technique, various kinds of models (e.g., decision trees, logistic regression, SVMs) are used together, and a meta-classifier is learned to learn how to combine the results of the individual models best. The stacking option permits models of varied strengths to supplement one another, which is especially applicable where fake news data are complex and heterogeneous.

Fake news is detected using ensemble models that have worked extensively in such a scenario. These not only increase the prediction quality but are also quite resistant to noisy and imbalanced data. Kaliyar et al. (2021) stressed that implementing ensemble learning in combination with NLP-based features showed better results than individual models.

Table 4: Hybrid Approaches in Fake News Detection.

Hybrid Approach	Components	Advantages	Example
ML/DL with NLP	NLP preprocessing + ML/DL classifier	Enhances feature representation and semantic understanding	CSI Model (Ruchansky et al., 2017)
Ensemble Learning	Multiple classifiers combined (Bagging, Boosting, Stacking)	Improves accuracy, reduces overfitting, and increases robustness	Random Forest, AdaBoost, Stacking with SVM and Logistic Regression

3. DATASETS USED IN FAKE NEWS DETECTION RESEARCH

Data is considered one of the most vital aspects of developing, training, and assessing fake news detecting models. Availability of quality datasets, labeled with some variance, allows researchers to check out different algorithms and confirm their suitability in real-life applications. Nonetheless, building up trustworthy datasets aimed at fake news detection is always challenging because of the problematic accuracy of labeling, representativeness, and the prejudice issue at hand. Several benchmark datasets were curated over the years and presented by different sources, each of which has contributed to the advancement of research in the field.

3.1 LIAR Dataset

Among the most popular benchmark datasets used in the studies of fake news detection is the LIAR dataset that was proposed by Wang (2017). It contains 12,836 online subtle statements of humans identified by politifact.com, one of the good websites/blogs where fact-checking is rendered. The labels given to each statement are finer-grained and have six categories: True, Mostly-True, Half-True, Barely-True, False, and Pants-on-Fire, instead of a binary label.

Besides the statements, the archive contains meta-information that includes the subject, the surrounding description, the speaker, party affiliation, and the speaker's history in terms of truthfulness. Models with enriched metadata that use textual content and contextual information can be explored. Regardless of its popularity, the main downside of the LIAR dataset is that it deals exclusively with short political statements that are not necessarily very complex compared to longer newspaper passages and different types of misinformation conditions.

3.2 Fake News Net Dataset

The Fake News Net dataset designed by Shu et al. (2018) is one of the complete databases seeking a multidimensional view of fake news detection. Contrary to the LIAR dataset, where emphasis is only made on the textual side of the data, Fake News Net considers the social context and the temporal information along with the contents. It collects news reports on fact-checking websites such as Politi Fact or Gossip Cop in both political and entertainment news spheres.

The uniqueness of Fake News Net is that it contains information on user engagement on social media platforms, e.g., likes, comments, and shares on platforms like Twitter. This renders it of great value to the research of fake news propagation, user behavior analysis, and the network-based ways of detecting fake news. Nevertheless, the fact that Fake News Net uses social media APIs to extract data complicates accessing and ensuring the consistency of the data when the privacy policies and the social media APIs are modified.

3.3 ISOT Fake News Dataset

The ISOT Fake News Dataset curated by Ahmed et al. (2017) has a large size, including 44,898 news articles that can be categorized into two groups: 21,417 honest and 23,481 fake news articles retrieved by the team of researchers on Reuters.com and different unreliable news websites accordingly. In contrast to the LIAR dataset, which consists of short claims, ISOT is composed of full-length news articles, and, therefore, it may be most appropriate when testing the models of fake news detection by content.

The ISOT dataset is helpful to test the effectiveness of the algorithms on text samples in a deeper domain since it is a more extensive sample that might be utilized in the context of deep learning, e.g., with CNNs and LSTMs. Nevertheless, the binary system of classifications (real or fake) can be too simplistic in explaining the nature of misinformation and lacks the grey nuances that are usually involved in such phenomena in real life.

3.4 Buzz Feed, Politi Fact, and Kaggle Datasets

Other data sets have been compiled with or without the help of a fact-checking body or obtained using open data repositories. Buzz Feed database Silverman et al. (2016) contains over 800 articles published on Facebook during the 2016 US presidential election; it is curated by journalists to be labeled as following the criterion of credibility. It gives us a firm understanding of political disinformation and the nature of social media.

On the same note, Politi Fact datasets, frequently paired with the LIAR dataset, provide sets of checked claims about different political aspects, and the resource is used in claim verification research. These datasets give more consideration to political fact-checking and can be adopted in research aiming at the perception of political misinformation.

Conversely, websites such as Kaggle have been the medium through which multiple challenges on fake news identification have been created, leading to the creation of standard datasets by the respective community, such as the Fake and Real News Dataset by George McIntire (2017). The data in these datasets are generally binary and are commonly used in machine learning challenges, where they become the sandbox used to test the algorithms. Nevertheless, the same source and cu ration may have altered their quality and labeling criteria.

3.5 Challenges with Dataset Quality and Bias

Although a few datasets are present, there remain a number of challenges that may impinge on the sensitivity and generalizability of results of fake news detection models. Labeling bias should be one of the most critical issues to address, i.e., most datasets are based on manual annotations

or third-party fact-checkers, which may lead to subjective judgment behind data classification. Also, the binary categorizing of news as fake or real simplifies the continuum of misinformation which usually covers the information to create satire, parodies, biased reporting, or partly accurate information.

The other serious issue is the representativeness of the sets of data. Most widely used datasets are gathered by Western media sources that mainly concentrate on English-speaking news. Thus, these datasets cannot be used to learn models with various applications in different languages or cultures. Moreover, the imbalance of the data (particularly, related to the overrepresentation of the classes of data, e.g., real news) may lead to biased performance of machine learning models, resulting in biased predictions.

Table 5: Commonly Used Fake News Datasets.

Dataset	Content Type	Size	Features	Limitations
LIAR	Short political claims	~12,800	Labels + Metadata (speaker info, context)	Focused on short claims, the political domain
FakeNews Net	News + Social media	Varies	Content + User engagement + Temporal data	Data accessibility via APIs
ISOT	Full news articles	~45,000	Binary labeled real/fake news articles	Simplified binary classification
BuzzFeed/ PolitiFact	Articles + Fact-checks	Varies	Political misinformation focus	Limited to specific events or timeframes
Kaggle Datasets	Mixed news articles	Varies	Community-generated, accessible datasets	Varied quality, simplified labels

4. EVALUATION METRICS

When it comes to fake news detection studies, it is essential to mention the evaluation of the classification models and their performance to determine their real-life applicability and dependability. Several evaluation metrics can be applied to measure the effectiveness of a fake news detection system in capturing the various aspects of model performance in terms of correctness, resilience, and tolerance to errors. These metrics help select the desired model compared to alternatives, and discuss the weaknesses and strengths. Nonetheless, the characteristics of fake news datasets, e.g., class imbalance and contextual variance, create specific problems in interpreting the given metrics.

4.1 Accuracy, Precision, Recall, and F1-Score

Among the most frequently reported values is the accuracy, which is the ratio of the adequately foreseen cases divided by the overall number of the forecasts made. Although accuracy gives a glimpse of the overall error rate of a model, it can be the worst of the models when used in situations where there is an imbalanced dataset, like the case with fake news detection, as the majority of the predictions are wrong in most of the prediction results (Sokolova & Lapalme, 2009).

Fewer black-and-white measures are often employed to overcome this, such as Precision, recall, and sometimes F1-score. Precision: The proportion of the predicted positives that are a positive outcome of the model is termed Precision. When dealing with fake news detection, Precision represents the accuracy by which the given model could detect the phony news without making a false alarm against real news as being fake. This is especially significant in the sensitive areas, where false positives may result in credibility concerns.

On the other hand, Recall measures the Precision of the actual positive cases that the model had correctly classified. High Recall would imply that the model can identify most fake news articles with few false negatives. As far as the task of misinformation detection is concerned, failing to recognize a false news item (false negative) can potentially lead to crucial effects on society, and thus, Recall is a critical measure (Goutte & Gaussier, 2005).

The harmonic above derives the F1-score as a fair compromise when a trade-off between Precision and Recall exists. F1-score is especially beneficial in the field of fake news detection studies since it uses only one value and considers both false positive and false negatives, giving a more holistic picture of the model's effectiveness under test in terms of imbalanced data.

4.2 ROC-AUC and Confusion Matrix

In addition to simple measurements, more comprehensive assessment instruments such as the Receiver Operating Characteristic (ROC) Curve and Area under the ROC Curve (AUC) are used. The ROC curve compares the false and actual favorable rates (sensitivity) at different threshold values. The AUC returns a single scalar number, which is a summary of model performance to differentiate between classes. The AUC value of 1.0 would mean that a model has an extreme discriminative power, and with an AUC of 0.5, this implies that the model is depicting no better performance than using

random guesses (Fawcett, 2006). AUC is beneficial in evaluating the comparison of models when the classification threshold is set at different levels, in fake news detection.

The second necessary tool is the confusion matrix, which details the model's performance in each class. It is comprised of four significant elements: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). By calculating the confusion matrix, one can see the nature of the model's errors directly; hence, target improvements can be made. For example, one may have many false positives (real news claimed to be fake), which could indicate that either the feature choice should be improved or decision thresholds need to be adjusted.

4.3 Limitations of Metrics in Fake News Detection

Although they will be helpful, the conventional evaluation measures are not good when dealing with fake news detection. To begin with, numerous phony news datasets are imbalanced, i.e., one class (real news) is outnumbered many times compared to the other. Here, some measures, such as accuracy, have an elusive perception about the model as a predictor that predicts that the majority can produce accuracy without identifying fake news (Haixiang et al., 2017).

Secondly, the standard measurement methods fail to include the varying values of damages when misclassified with classification errors of different types. An example is the fact that successfully identifying a fake news article that is very misleading (false negative) may have more severe sweeping effects on society than a false positive of labeling an actual news story. Hence, recognizing the application context should explain the Precision, Recall, and F1-score measures.

Table 6: Evaluation Metrics in Fake News Detection.

Metric	Definition	Importance of Fake News Detection	Limitation
Accuracy	Correct predictions / Total predictions	Provides an overall performance snapshot	Misleading with imbalanced datasets
Precision	TP / (TP + FP)	Measures the correctness of positive (fake news) predictions	May sacrifice recall
Recall (Sensitivity)	TP / (TP + FN)	Measures the ability to detect all actual fake news	May sacrifice Precision
F1-Score	Harmonic mean of Precision and Recall	Balances Precision and recall in a single metric	Sensitive to data imbalance
ROC-AUC	Area under the ROC curve	Evaluates discriminative power across thresholds	Requires probabilistic outputs
Confusion Matrix	Tabular representation of TP, TN, FP, FN	Offers granular insight into model performance	Does not offer a summary performance measure

5. CHALLENGES AND RESEARCH GAPS

Even though the research in fake news detection has progressed tremendously, many challenges and gaps remain that drag on reaching a reliable and robust fake news detection solution. The constraints in access to data, model structure, multicultural language, and dynamicity in misinformation cause such difficulties. These issues are to be addressed so that automated instruments of fake news detection become more effective and applicable to real-life contexts.

5.1 Data Scarcity and Imbalance

Limitations and insufficiencies of high-quality, labeled datasets are among the most urgent problems related to fake news detection. Creation of comprehensive datasets demands the manual labeling of the dataset by domain experts or fact-checkers, which is resource-intensive and time-consuming. Therefore, the majority of the currently available data, like LIAR (Wang, 2017) or ISOT (Ahmed et al., 2017), is small, confined, or domain-specific. Such a limitation impedes training data-hungry, intense learning models that require vast amounts of data to achieve optimal performance.

Besides scarcity, the frequent problem is the balance in the classes. Naturally, the information ecosystem contains a bias in the ratio of real news to fake news, which is similarly indicated in datasets. Such an imbalance results in biased models that might attain a high score but are limited in recognizing fake news since it is biased towards the higher class (Haixiang et al., 2017). Resampling, synthetic data generation or cost-sensitive learning are the methods that have been considered; in most cases, they bring about issues such as over fitting or loss of data diversity.

5.2 Generalization across Platforms and Languages

The other research gap that can be described as critical is the generalization ability of the fake news detection models. The characteristic feature of most existing research works is that they prepare and test the models on data obtained using a particular platform, e.g., Twitter, Facebook, or certain news media, which restricts their practical usage on other media

environments. For example, a model developed using Twitter data might not work on long news stories of the online sources because of variance in the text structure, the use of languages, and the context (Shu et al., 2020).

Additionally, the findings of most work in fake news detection have been in the English language, thus leaving a high gap to fill in misinformation among people not well versed in the language. The global misinformation has a lot of linguistic diversity that requires adopting models capable of dealing with multilingual datasets and cross-lingual transfer. Nonetheless, creating such models is associated with issues concerning the specificities of the language, cultural background, and diverse language peculiarities.

5.3 Model Explain ability and Transparency

The inability to explain and be transparent about the fake news detection models, especially those utilizing deep learning, is a significant obstacle to the uptake and reliability of the models. Such models as CNNs, LSTMs, and Transformers are opaque, and thus their users, developers, or regulators cannot comprehend the decision-making mechanism (Doshi-Velez and Kim, 2017). Generating predictions that are not only accurate but also intelligible is paramount to the stakeholders in cases where they need to use predictions, such as in content moderation, journalism, or law enforcement.

There has been a proposal of explainable AI (XAI) frameworks, e.g., LIME (local interpretable model-agnostic explanations) or SHAP (Shapley Additive explanations), to offer post-hoc explanations about the model decision. Nevertheless, these approaches usually have limited applications in time or computing costs, or are too uninformative to draw valuable conclusions in other linguistic situations. The question that still needs to be addressed lies in designing models that are very precise and interpretable, as the systems working in automated fake news detection are more accountable and transparent.

5.4 Adversarial Attacks and Misinformation Evolution

Another significant complication consists of the dynamic character of misinformation. The misinformation distributors never stop changing their tactics, and they make their fake news more complex, less noticeable, and contextually appropriate to avoid detection mechanisms. The concept drift issue has been determined, wherein the models designed on the historical information do not identify the new trend of fake news (Shu et al., 2020). Conventional static models do not cope quickly with such a pace of change, with frequent retraining and updating of these models.

Also, adversarial attacks specifically against fake news detection systems have become more of a threat as the malicious actors try to manipulate the text input so that they can deceive the classifiers. It is possible to remarkably reduce the performance of models with techniques such as paraphrasing, inserting irrelevant words, or synonym replacement (Zhou et al., 2020). It has been observed that current models are usually not robust when targeted via adversarial manipulations. Thus, developing more robust architectures that can survive deliberate evasion tactics becomes imperative.

6. CONCLUSIONS

Fake news detection has become one of the topical directions of research in the face of threats in the digital era, given the danger of misinformation. This review has reviewed a variety of solutions taken in fake news detection, including the application of the rule-based systems and the usage of keywords at the beginning of their development, to more elaborate systems based on machine learning (ML) and deep learning (DL) systems, coupled with natural language processing (NLP) systems. Simple systems, such as early systems, were applied to the automation foundation using handmade rules and manual feature extraction. Nevertheless, these methods were challenged by usability and flexibility concerning the linguistic patterns of misinformation, which are intricate.

Implementing machine learning models and especially supervised learning approaches such as Naive Bayes, Support Vector Machines, Logistic Regression, and Random Forest fooled many scientists by unlocking statistical characteristics in data. When used together with feature engineering techniques like TF-IDF, Bag-of-Words, and n-grams based on NLP, these models increased detection accuracy since they centered on analyzing textual content. Unsupervised learning styles also helped by discovering hidden data structures in clustering and topic modeling, but with fair results in free-form classification.

Nevertheless, despite these gains, the analysis has highlighted the need for hybrid approaches wherein strengths will be combined through different methods. Incorporating NLP and ML/DL models increases the representatively of features and learning abilities of systems and allows them to process different linguistic phenomena and variations in content. In addition, ensemble learning models and hybrid frameworks have demonstrated their potential to improve performance indicators and address such issues as class imbalance or adversarial manipulation.

REFERENCES

1. H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking," in EMNLP, pp. 2931–2937, 2017. DOI: 10.18653/v1/D17-1318

2. N. Ruchansky, S. Seo, and Y. Liu, "CSI: A Hybrid Deep Model for Fake News Detection," in CIKM, pp. 797–806, 2017. DOI: 10.1145/3132847.3132877

3. K. Shu, S. Wang, and H. Liu, "Fake News Detection on Social Media: A Data Mining Perspective," ACM SIGKDD Explor. Newsl., vol. 19, no. 1, pp. 22–36, 2017. DOI: 10.1145/3137597.3137600

4. W. Wang, "'Liar, Liar Pants on Fire': A New Benchmark Dataset for Fake News Detection," in ACL, pp. 422–426, 2017. DOI: 10.18653/v1/P17-2067

5. V. Perez-Rosas, B. Kleinberg, A. Lefevre and R. Mihalcea, "Automatic Detection of Fake News," in COLING, pp. 3391–3401, 2018. DOI: 10.18653/v1/C18-1288

6. H. Ahmed, I. Traore and S. Saad, "Detecting Opinion Spams and Fake News Using Text Classification," Security and Privacy, vol. 1, no. 1, e9, 2017. DOI: 10.1002/spy2.9

7. A. Bondielli and F. Marcelloni, "A Survey on Fake News and Rumour Detection Techniques," Inf. Sci., vol. 497, pp. 38–55, 2019. DOI: 10.1016/j.ins.2019.05.035

8. R. Kaliyar, A. Goswami and P. Narang, "FakeBERT: Fake News Detection in Social Media with a BERT-Based Deep Learning Approach," Multimed. Tools Appl., vol. 80, pp. 11765–11788, 2021. DOI: 10.1007/s11042-020-10183-2

9. A. Oshikawa, J. Qian and W. Y. Wang, "A Survey on Natural Language Processing for Fake News Detection," in Proc. LREC, pp. ××, 2020.

10. K. Shu et al., "FakeNewsNet: A Data Repository with News Content, Social Context and Spatiotemporal Information for Studying Fake News on Social Media," Big Data, vol. 8, no. 3, pp. 171–188, 2020. DOI: 10.1089/big.2020.0066

11. C. Silverman et al., "Hyperpartisan Facebook Pages Are Publishing False and Misleading Information at an Alarming Rate," BuzzFeed News, Dec. 2016.

12. G. McIntire, "Fake and Real News Dataset," Kaggle, 2017.

13. K. Zhang and A. A. Ghorbani, "An Overview of Online Fake News: Characterization, Detection, and Discussion," Inf. Process. Manag., vol. 57, no. 2, p. 102025, 2020. DOI: 10.1016/j.ipm.2019.102025

14. Ray Oshikawa, Jing Qian, and William Y. Wang, "A Survey on Natural Language Processing for Fake News Detection," arXiv, Nov. 2018.

15. H. Zhou, J. Ebrahimi and D. Dou, "Defense Against Adversarial Attacks in NLP: Methods and Challenges," in Proc. ACL, pp. 4719–4731, 2020.

16. K. Goldani, S. Momtazi and R. Safabakhsh, "Detecting Fake News with Capsule Neural Networks," arXiv, Feb. 2020.

17. J. Y. Khan et al., "A Benchmark Study of Machine Learning Models for Online Fake News Detection," arXiv, May 2019.

18. M. Choras et al., "Advanced Machine Learning Techniques for Fake News (Online Disinformation) Detection: A Systematic Mapping Study," arXiv, Jan. 2021.

19. K. Roumeliotis, N. Tselikas and D. Nasiopoulos, "Fake News Detection and Classification: A Comparative Study of CNNs, Large Language Models, and NLP Models," Future Internet, vol. 17, no. 1, 28, 2025. DOI: 10.3390/fi17010028

20. S. Shah and S. Patel, "A Comprehensive Survey on Fake News Detection Using Machine Learning," J. Comput. Sci., vol. 21, no. 4, pp. 982–990, 2025. DOI: 10.3844/jcssp.2025.982.990

21. P. Swetha and E. Priyanka, "Fake News Detection on Social Media Using Regional CNN-RNN and BERT," J. Comput. Soc. Media, 2023.

22. Daksha Tiwari and S. K. Sharma, "Sentiment Aware Fake News Detection on Online Social Networks," Proc. IEEE ICPCA, pp. ××–××, 2019.

23. O. Ajao, D. Bhowmik and S. Zargari, "Sentiment Aware Fake News Detection on Online Social Networks," in Proc. ACM, 2019.

24. S. Nakov et al., "Predicting Factuality of Reporting and Bias of News Media Sources," ACL, Nov. 2018.

25. Kai Shu et al., "dEFEND: Explainable Fake News Detection," in Proc. ACM, Jul. 2019.

26. J. Wang et al., "Large Language Model Based Fake News Detection," J. Inf. Tech. Soc., 2023.

27. D. Khattar et al., "MVAE: Multimodal Variational Autoencoder for Fake News Detection," in Proc. ACM, May 2019.

28. M. Ebrahimi and A. H. Dou, "Detecting fake news by exploring consistency of multimodal data," PLoS One, 2022. DOI: 10.1371/journal.pone.0264561

29. K. Conroy, V. Rubin and Y. Chen, "Automatic Deception Detection: Methods for Finding Fake News," Assoc. Inf. Sci. Tech., vol. 52, no. 1, pp. 1–4, 2015.

30. T. Fawcett, "An Introduction to ROC Analysis," Pattern Recognit. Lett., vol. 27, no. 8, pp. 861–874, 2006. DOI: 10.1016/j.patrec.2005.10.010