



“A Data-Driven Machine Learning Framework for Identifying Misinformation in Online News “

Dr. Satyendra Kumar Bunkar, Associate Professor, Dept: Computer Science and Engineering,
Chameli Devi Group of Institutions, Indore, satyendra.bunkar@gmail.com

How to Cite this Article:

Bunkar, S. K. (2026). A Data-Driven Machine Learning Framework for Identifying Misinformation in Online News. International Journal of Creative and Open Research in Engineering and Management, <i>02</i>(9), 1-9. <https://doi.org/10.55041/ijcope.v2i9.229>

License:

This article is published under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

© The Author(s). Published by International Journal of Creative and Open Research in Engineering and Management.



<https://doi.org/10.55041/ijcope.v2i9.229>

Abstract

The rapid growth of online news platforms and social media has significantly increased the dissemination of misinformation, creating challenges for individuals, organizations, and digital platforms in distinguishing credible information from misleading content. This study proposes a data-driven machine learning framework for identifying misinformation in online news by leveraging natural language processing and supervised machine learning techniques. The proposed framework involves data pre-processing, text normalization, feature extraction, model training, classification, and performance evaluation. Various textual features, including term frequency, word importance, linguistic patterns, and semantic representations, are extracted from online news articles to identify characteristics associated with misinformation. Multiple machine learning algorithms, including Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), and Gradient Boosting, is considered to evaluate their effectiveness in classifying news articles as reliable or misleading. The models are evaluated using standard performance metrics such as accuracy, precision, recall, F1-score, and confusion matrix analysis. In addition, feature importance analysis is incorporated to identify the textual characteristics that contribute significantly to misinformation detection. The proposed data-driven

approach aims to provide an automated and scalable mechanism for detecting potentially misleading online news while supporting transparency in the classification process. The findings of this research can contribute to the development of intelligent misinformation-monitoring systems for digital news platforms, social media applications, and information-verification environments.

Keywords— Misinformation Detection, Machine Learning, Natural Language Processing, Online News, Fake News Detection, Text Classification, Support Vector Machine, Random Forest, Feature Extraction.

1. INTRODUCTION

In recent years, the rapid expansion of digital communication and online news platforms has transformed the way people access, consume, and share information. News can now be distributed globally within seconds through websites, social media platforms, blogs, and online communities. While this transformation has improved the accessibility and speed of information dissemination, it has also created favourable conditions for the rapid spread of misinformation. Misleading or inaccurate information can influence public perception, create confusion, damage the credibility of legitimate information sources, and produce significant social and economic consequences. Therefore, developing effective and automated techniques for identifying misinformation in online news has become an important research area.

The increasing dependence on digital media has made online news an essential source of information for individuals, organizations, and institutions. Unlike traditional media, online platforms allow users not only to consume information



but also to generate, modify, and redistribute content. The ease of sharing information enables a news article to reach a large audience in a very short period. However, the same characteristics can also facilitate the dissemination of inaccurate, misleading, or fabricated information. Consequently, distinguishing trustworthy news from misinformation has become increasingly difficult, particularly when large volumes of online content are generated continuously.

Misinformation can appear in several forms, including misleading headlines, manipulated narratives, and inaccurate claims, fabricated news articles, selectively presented facts, and misleading textual content. It may be generated or distributed for different purposes, such as obtaining financial benefits, increasing website traffic, influencing public opinion, promoting particular interests, or creating confusion. In addition, misinformation may not always be completely false; it can contain a combination of factual and misleading information, making its identification more challenging. These characteristics create significant difficulties for conventional manual fact-checking approaches, which can be time-consuming and difficult to scale.

The identification of misinformation is therefore important for maintaining the reliability and integrity of digital information ecosystems. Continuous exposure to misleading information can affect how users understand events and make decisions. The increasing volume of online news also makes it impractical to manually verify every article. An automated computational approach capable of analyzing large collections of news articles and identifying linguistic and statistical patterns associated with misinformation can assist researchers, journalists, content moderators, and digital platforms in the preliminary identification of potentially misleading content.

Machine learning (ML) provides an effective approach for addressing this problem because ML algorithms can learn patterns from previously labelled data and apply the learned patterns to previously unseen instances. In the context of misinformation detection, machine learning models can analyze textual characteristics of news articles and classify them according to their learned patterns. Different algorithms, such as Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest, Naïve Bayes, and Gradient Boosting, can be applied to investigate the relationship between textual features and misinformation classes.

Natural Language Processing (NLP) plays an important role in transforming unstructured news content into machine-processable representations. Text pre-processing techniques such as tokenization, removal of stop words, stemming or lemmatization, and normalization can be applied to improve the quality of the input data. Subsequently, feature extraction techniques such as Term Frequency–Inverse Document Frequency (TF-IDF), Bag-of-Words, and word embeddings can be employed to represent textual information numerically. These representations enable machine learning algorithms to identify vocabulary patterns, word associations, and other characteristics that may contribute to the classification of online news.

One of the major challenges in misinformation detection is the complexity and dynamic nature of online information. The language used in news articles changes over time, and misinformation can be presented using sophisticated writing styles that resemble legitimate journalism. Furthermore, differences in topics, writing styles, sources, datasets, and social contexts can affect the performance of machine learning models. A model trained on a particular dataset may not necessarily perform similarly when applied to news from a different domain or time period. Therefore, a data-driven framework should consider appropriate pre-processing, feature engineering, model selection, and systematic evaluation to improve the reliability of misinformation identification.

The proposed research, “A Data-Driven Machine Learning Framework for Identifying Misinformation in Online News,” focuses on developing a systematic machine learning-based approach for classifying online news content. The framework considers the complete processing pipeline, including dataset collection, data pre-processing, and feature extraction, machine learning model development, classification, and performance evaluation. Multiple machine learning algorithms can be comparatively investigated using evaluation measures such as accuracy, precision, recall, F1-score, and confusion matrix. Feature importance analysis can additionally be used to investigate which textual characteristics contribute to the classification process.



The overall objective of this research is to develop a scalable and data-driven approach that can assist in identifying potentially misleading information in online news. Rather than relying exclusively on manual verification, the proposed approach aims to use computational learning techniques to analyze large volumes of textual data efficiently. Such a framework can contribute to the development of intelligent misinformation-monitoring and information-verification systems for online news platforms, social media environments, and other digital information ecosystems.

2. Research Gap

The literature indicates that existing studies have successfully applied machine learning, deep learning, NLP, ensemble learning, and social-context analysis to misinformation detection. However, many approaches focus primarily on achieving classification performance on individual benchmark datasets. The literature also highlights the importance of dataset characteristics, feature representation, model selection, explain ability, and domain generalization.

Therefore, there is scope for a systematic data-driven framework that integrates the complete pipeline of misinformation identification, including data pre-processing, feature extraction, comparative machine learning model development, performance evaluation, and feature-importance analysis. The proposed research addresses this gap by developing a framework for identifying misinformation in online news using multiple machine learning algorithms and standardized evaluation measures. Such an approach can facilitate a more systematic comparison of models and provide insights into the textual features that contribute to misinformation classification.

3. LITERATURE REVIEW

Pérez-Rosas et al. [5] investigated the automatic detection of fake news using linguistic characteristics of online news articles. The researchers introduced two datasets covering multiple news domains and examined differences between fake and legitimate news content. They applied machine learning techniques to develop automated fake-news classifiers and reported accuracies of up to 76%. Their findings demonstrated that linguistic characteristics can provide useful signals for distinguishing misleading content from legitimate news. The study established an important foundation for data-driven fake-news detection by demonstrating the feasibility of combining linguistic analysis with supervised machine learning.

Zhou and Zafarani [6] presented a comprehensive survey of fake-news detection methods and categorized existing approaches from four major perspectives: knowledge, writing style, propagation patterns, and source credibility. Their study highlighted that misinformation detection is not limited to analyzing the textual content of an article; information about how content spreads and the credibility of its source can also provide important detection signals. The authors further discussed the need for efficient and explainable detection techniques and identified several research opportunities for developing more reliable misinformation-detection systems.

Ahmad et al. [7] proposed a machine learning ensemble approach for automated fake-news classification. Their research investigated different textual properties that could distinguish fake content from authentic news and combined multiple machine learning classifiers using ensemble-learning strategies. The proposed approach was evaluated on four real-world datasets, demonstrating the potential of combining several classifiers instead of relying on a single machine learning algorithm. The study indicates that ensemble methods can provide a useful framework for improving the robustness of automated misinformation classification.

Shu et al. [8] introduced DEFEND, an explainable fake-news detection framework designed to address an important limitation of conventional classification systems: the inability to explain why a particular article is classified as fake. The proposed framework jointly considers news content and user comments through a sentence-comment co-attention mechanism. The model identifies important sentences and comments that contribute to the prediction, thereby providing explanations in addition to classification. Experimental results on real-world datasets demonstrated improvements over several existing approaches and showed that the explanations could help users understand the



model's decisions. This work emphasizes the importance of interpretability and transparency in automated misinformation detection.

Wani et al. [9] investigated deep-learning approaches for detecting COVID-19-related fake news. The researchers evaluated supervised text-classification approaches based on CNN, LSTM, and BERT architectures using a COVID-19 fake-news dataset. They also investigated the contribution of language-model pre-training and distributed word representations. Their experimental results showed that deep learning and transformer-based representations can be effective for misinformation detection in domain-specific datasets. The study also demonstrates that model performance can be strongly influenced by the characteristics and domain of the dataset being analyzed.

Zhou et al. [10] proposed a theory-driven approach for early fake-news detection, addressing the limitation of methods that depend on social-media propagation information. The authors noted that propagation-based features may not be available immediately after a news article is published. Their work therefore investigated the possibility of identifying fake news at an early stage, before substantial dissemination occurs. This research is particularly relevant to data-driven misinformation detection because it highlights the importance of developing models that can operate using information available at the initial publication stage.

Research on fake-news datasets has also emphasized the importance of dataset quality and evaluation methodology. Surveys of fake-news detection datasets have shown considerable variation in dataset characteristics, including data sources, labelling procedures, topics, languages, and available contextual information. Such variations can influence the generalization capability of machine learning models. Consequently, the construction and selection of appropriate benchmark datasets are critical components of a reliable misinformation-detection framework.

Zhou and Maharani [11] further emphasized that fake-news detection can be approached through content-based, propagation-based, source-based, and hybrid techniques. Content-based methods primarily examine the information contained within news articles, whereas propagation-based approaches analyze how information spreads through networks. Source-based approaches consider the credibility and characteristics of information sources. The authors highlighted that combining complementary information sources may provide opportunities for developing more comprehensive detection systems.

Shu et al. [12] demonstrated that incorporating social context can improve the interpretability of fake-news detection. Their work illustrates that news content alone may not always provide sufficient evidence for determining the reliability of an article. User comments and related social signals can provide additional information that helps identify check-worthy content. This suggests that future data-driven frameworks can benefit from integrating textual and contextual information rather than relying exclusively on conventional word-frequency features.

Existing research also demonstrates the importance of combining feature extraction and machine learning algorithms. Traditional approaches commonly employ features such as Bag-of-Words and TF-IDF representations before applying classifiers such as Logistic Regression, Support Vector Machines, Naïve Bayes, Random Forest, and other ensemble algorithms. More recent approaches increasingly employ distributed word representations and transformer-based language models. The progression from manually engineered textual features toward contextual representations indicates a continuing shift toward models capable of capturing richer linguistic information.

Despite significant progress, several challenges remain in automated misinformation detection. First, the performance of machine learning models can vary considerably across datasets and domains. Second, misinformation can be expressed using language that is structurally similar to legitimate journalism, making purely text-based classification difficult. Third, models trained on historical datasets may experience reduced performance when applied to newly emerging topics and changing linguistic patterns. Finally, highly accurate models may still have limited practical value if their decisions cannot be interpreted or verified by human users.



4. RESEARCH METHODOLOGY

The proposed research presents a data-driven machine learning framework for identifying misinformation in online news. The framework utilizes Natural Language Processing (NLP) techniques to transform unstructured news articles into meaningful numerical representations and applies supervised machine learning algorithms to classify news content as reliable or potentially misleading. The proposed methodology consists of six major stages: (1) data collection, (2) data pre-processing, (3) feature extraction and representation, (4) machine learning model development, (5) model evaluation, and (6) comparative analysis and interpretation.

The overall objective is to develop a systematic classification pipeline capable of learning patterns from labeled news data and applying the learned patterns to previously unseen news articles. Multiple machine learning models are considered rather than depending on a single algorithm. This allows the performance of different learning approaches to be compared under the same experimental conditions.

A. Proposed System Design

The proposed framework consists of the following components:

- **Input Data:** Collection of labelled online news articles containing reliable and misinformation instances.
- **Data Cleaning:** Removal of irrelevant, incomplete, duplicate, and noisy records.
- **Text Pre-processing:** Conversion of text into a standardized form through normalization, tokenization, stop-word removal, and lemmatization.
- **Feature Extraction:** Transformation of textual information into numerical features using techniques such as TF-IDF and n-gram representation.
- **Feature Selection:** Identification of relevant features that contribute to the classification process.
- **Model Training:** Training of different supervised machine learning algorithms using the extracted features.
- **Classification:** Prediction of the class of previously unseen news articles.
- **Performance Evaluation:** Evaluation using accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix.
- **Feature Importance Analysis:** Identification of important textual characteristics associated with the classification decision.
- **Comparative Analysis:** Comparison of different machine learning models to understand their behaviour under the same experimental conditions.

The proposed system can be represented by the following workflow:

Online News Dataset → Data Cleaning → Text Pre-processing → Feature Extraction → Feature Selection → Train/Test Split → ML Model Training → Prediction → Performance Evaluation → Feature Importance Analysis

B. Data Collection

Data collection is the first stage of the proposed methodology. A suitable dataset containing both reliable and misinformation-labelled online news articles is required for supervised learning. Publicly available benchmark datasets may be utilized to ensure reproducibility and facilitate comparison with previous research.

The dataset should preferably contain fields such as:

- News article title
- Article body/text
- Subject/category
- Publication information, where available



- Source information, where available
- Class/label indicating reliable or misinformation content

For the proposed research, the title and article body can be combined to construct the primary textual input. Records containing missing text, duplicate articles, or invalid labels are removed during data preparation.

The collected dataset is divided into two classes:

$$Y \in \{0, 1\} \quad Y \in \{0, 1\}$$

Where:

- $Y=0, Y=0$ represents reliable news.
- $Y=1, Y=1$ represents misinformation.

The distribution of the two classes is analyzed before model development. If substantial class imbalance is observed, appropriate techniques such as stratified sampling or class weighting can be considered.

C. Data Pre-processing

Raw online news data generally contains noise, inconsistencies, HTML elements, punctuation, URLs, special symbols, and other information that may not contribute directly to textual classification. Therefore, pre-processing is performed before feature extraction.

The following pre-processing operations are applied:

1. Missing-Value Removal

Records with missing or insufficient textual information are identified. Articles that do not contain meaningful text are removed from the dataset.

2. Duplicate Removal

Duplicate articles are detected and removed to prevent the same information from appearing in both training and testing datasets. This is particularly important because duplicate or near-duplicate articles can lead to artificially high classification performance.

3. Text Normalization

The textual content is converted into a standardized representation. This includes:

- Conversion of text to lowercase
- Removal of unnecessary whitespace
- Removal of HTML tags
- Removal of URLs where appropriate
- Removal of special characters
- Removal of unnecessary punctuation

4. Tokenization

The cleaned text is divided into individual tokens or words. For example:



Input:

“The government announced a new policy today.”

After tokenization:

[government, announced, new, policy, today]

5. Stop-Word Removal

Frequently occurring words that provide limited discriminative information, such as *the*, *is*, *are*, *of*, and *and*, may be removed. However, stop-word removal should be experimentally validated because some function words can contain useful stylistic information.

6. Lemmatization

Words are converted into their basic dictionary forms. For example:

“running”, “runs”, “ran” → “run”

Lemmatization reduces unnecessary variation in vocabulary while retaining the semantic identity of words.

D. Feature Extraction and Text Representation

After pre-processing, the textual information must be transformed into numerical features that can be processed by machine learning algorithms.

1. TF-IDF Representation

The primary feature extraction technique used in the proposed framework is Term Frequency–Inverse

For a term t in document d , the TF-IDF weight is defined as:

$$TFIDF(t, d) = TF(t, d) \times IDF(t)$$

where the Term Frequency (TF) can be calculated as:

$$TF(t, d) = \frac{f(t, d)}{\sum_{t' \in d} f(t', d)}$$

Here, $f(t, d)$ represents the number of times term t occurs in document d .

The Inverse Document Frequency (IDF) is defined as:

$$IDF(t) = \log \left(\frac{N}{df(t) + 1} \right)$$



Here:

- NN = total number of documents
- $df(t)df(t)$ = number of documents containing term tt
- $TF(t,d)TF(t,d)$ = frequency of term tt in document dd

TF-IDF assigns higher importance to words that are frequent in a particular document but relatively uncommon across the complete collection.

2. N-Gram Features

Unigrams and bigrams can be considered to capture both individual words and short sequences of words.

For example:

Unigram:

“false”, “claim”, “government”

Bigram:

“false claim”, “government announced”

N-gram features can help capture contextual expressions that may not be represented adequately by individual words.

3. Optional Semantic Representation

To extend the framework, word or sentence embeddings can additionally be investigated. Pre-trained representations such as Word2Vec, Glove, or transformer-based embeddings can be used in a separate experimental setting.

This enables comparison between traditional statistical text representation and semantic representation.

E. Feature Selection The dimensionality of textual datasets can become very large because thousands of unique words and n-grams may be present. Therefore, feature selection can be applied to identify informative features.

Possible techniques include:

- Chi-square feature selection
- Mutual information
- Variance-based filtering
- Select K Best
- Model-based feature importance

Feature selection can reduce computational complexity and potentially improve generalization by removing irrelevant or redundant features.

F. Machine Learning Model Selection

To evaluate the effectiveness of different machine learning approaches, the proposed framework considers multiple supervised classification algorithms.



1. Logistic Regression

Logistic Regression is used as a baseline classification model for binary classification. It estimates the probability that a given news article belongs to the misinformation class.

The logistic function is:

$$P(Y = 1 | X) = \frac{1}{1 + e^{-z}}$$

where

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

Therefore, the complete Logistic Regression equation can be written as:

$$P(Y = 1 | X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

Logistic Regression is particularly suitable for high-dimensional sparse textual features such as TF-IDF.

2. Support Vector Machine

Support Vector Machine (SVM) attempts to identify an optimal decision boundary that separates the two

If you are describing a Support Vector Machine (SVM), you can write:

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$$

The classification rule is:

$$\hat{y} = \begin{cases} +1, & f(\mathbf{x}) \geq 0 \\ -1, & f(\mathbf{x}) < 0 \end{cases}$$

For your misinformation-detection problem, you may define:

- $+1$ = Misinformation
- -1 = Reliable News

The corresponding decision boundary is:

$$\mathbf{w}^T \mathbf{x} + b = 0$$

3. Random Forest

Random Forest is an ensemble learning algorithm consisting of multiple decision trees. Each tree generates a prediction, and the final classification is obtained through aggregation of the individual tree predictions.



The model can also provide feature-importance information, which is useful for understanding which textual characteristics contribute to the classification process.

4. Naïve Bayes

Naïve Bayes is a probabilistic classification algorithm commonly applied to text-classification tasks. It assumes conditional independence among features given the class.

Given an input feature vector $X = (x_1, x_2, \dots, x_n)$, the posterior probability of class C is given by Bayes' theorem:

$$P(C | X) = \frac{P(C) P(X | C)}{P(X)}$$

Under the Naïve Bayes conditional independence assumption:

$$P(X | C) = \prod_{i=1}^n P(x_i | C)$$

Its computational simplicity makes it useful as a baseline model.

5. Gradient Boosting

Gradient Boosting can also be incorporated to investigate whether sequential ensemble learning improves classification performance. The model constructs a series of weak learners, with each subsequent learner attempting to reduce the errors produced by previous learners.

G. Proposed Model Training

After feature extraction, the dataset is divided into training and testing subsets. A stratified split is recommended so that the proportions of reliable and misinformation articles remain approximately consistent in both subsets.

For example:

- 80% Training Data
- 20% Testing Data

The training dataset is used to learn the relationship between textual features and class labels, while the testing dataset is kept separate for final evaluation.

An additional validation strategy, such as k-fold cross-validation, can be employed during model development. This provides a more reliable estimate of model behaviour and reduces dependence on a single train-test partition.

The general training process is:

Training Dataset → TF-IDF Transformation → Feature Selection → ML Algorithm → Learned Classification Model

The same pre-processing and feature-transformation parameters learned from the training data must be applied to validation and testing data to avoid data leakage.



H. Hyper Parameter Optimization: The performance of machine learning algorithms can depend significantly on their hyper parameters. Therefore, hyper parameter optimization can be incorporated into the proposed framework.

Possible methods include:

- Grid Search
- Randomized Search
- Cross-validation-based optimization

Examples of parameters include:

Algorithm	Important Parameters
Logistic Regression	C, solver, class weight
SVM	C, kernel, class weight
Random Forest	Number of trees, maximum depth, minimum samples
Naïve Bayes	Smoothing parameter
Gradient Boosting	Number of estimators, learning rate, maximum depth

The optimization process should be performed using only the training data or cross-validation folds, while the final test set remains untouched until the final evaluation.

I. Model Evaluation

The trained models are evaluated using multiple performance measures rather than relying only on accuracy.

1. Accuracy

Accuracy measures the proportion of correctly classified instances.

$$Precision(\%) = \frac{TP}{TP + FP} \times 100$$

where:

- TP = True Positive
- TN = True Negative
- FP = False Positive
- FN = False Negative

2. Precision

Precision measures how many articles classified as misinformation are actually misinformation.



$$Precision = \frac{TP}{TP + FP}$$

Where:

- TP = Misinformation articles correctly identified as misinformation
- FP = Reliable articles incorrectly classified as misinformation

The precision percentage is:

$$Precision(\%) = \frac{TP}{TP + FP} \times 100$$

3. Recall

Recall measures how many actual misinformation articles are correctly identified.

$$Recall = \frac{TP}{TP + FN}$$

Where:

- TP = Misinformation articles correctly identified as misinformation
- FN = Misinformation articles incorrectly classified as reliable

The recall percentage is:

$$Recall(\%) = \frac{TP}{TP + FN} \times 100$$

4. F1-Score

F1-score provides a harmonic mean of precision and recall.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

5. ROC-AUC

The Receiver Operating Characteristic–Area Under the Curve (ROC-AUC) can be used to evaluate the model's ability to distinguish between the two classes across different classification thresholds.

6. Confusion Matrix

A confusion matrix provides a detailed representation of:



- True Positives
- True Negatives
- False Positives
- False Negatives

This is particularly important for misinformation detection because false classifications can have different implications depending on the application.

J. Feature Importance and Interpretability

An important component of the proposed data-driven framework is the analysis of the features contributing to model predictions.

For interpretable models such as Logistic Regression, feature coefficients can be examined to identify terms associated with each class. Random Forest and other tree-based models can provide feature-importance measures.

For more complex models, model-agnostic techniques such as SHAP (Shapley Additive explanations) can be considered to explain individual predictions and identify influential features.

The feature-importance analysis can answer questions such as:

- Which words contribute strongly to misinformation classification?
- Which textual patterns are associated with reliable news?
- Are particular n-grams disproportionately important?
- Does the model rely on meaningful linguistic patterns or potentially spurious features?

This component increases the transparency of the proposed framework.

K. Comparative Analysis

After training all selected models, their performance is compared using the same test dataset and evaluation criteria.

A comparative table can be constructed as follows:

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	To be calculated	To be calculated	To be calculated	To be calculated	To be calculated
Naïve Bayes	To be calculated	To be calculated	To be calculated	To be calculated	To be calculated
SVM	To be calculated	To be calculated	To be calculated	To be calculated	To be calculated
Random Forest	To be calculated	To be calculated	To be calculated	To be calculated	To be calculated
Gradient Boosting	To be calculated	To be calculated	To be calculated	To be calculated	To be calculated

The experimental results will be used to understand how different machine learning approaches behave when applied to the same misinformation-detection task.



L. Proposed Algorithm

Algorithm 1: Data-Driven Misinformation Detection

Input: Labelled online news dataset DD

Output: Predicted class YY

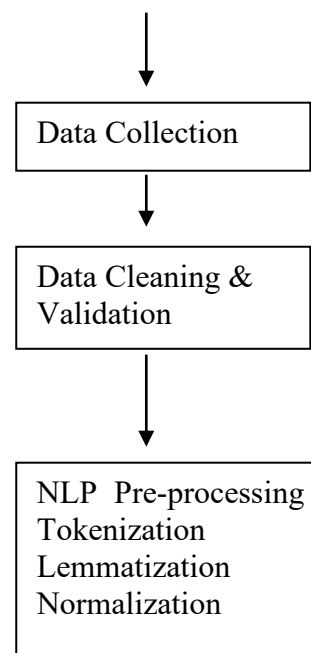
1. Collect labelled online news articles.
2. Remove duplicate and invalid records.
3. Handle missing textual values.
4. Normalize and clean the news text.
5. Tokenize the textual content.
6. Remove or evaluate stop words.
7. Apply lemmatization.
8. Divide the dataset into training and testing subsets using stratification.
9. Generate TF-IDF and n-gram features using the training data.
10. Apply feature-selection techniques where required.
11. Train Logistic Regression, Naïve Bayes, SVM, Random Forest, and Gradient Boosting models.
12. Optimize model hyper parameters using cross-validation.
13. Apply the trained models to the unseen test dataset.
14. Generate predicted class labels.
15. Calculate accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix.
16. Perform feature-importance and explain ability analysis.
17. Compare the performance of all models.
18. Analyze the results and identify limitations and future improvements.

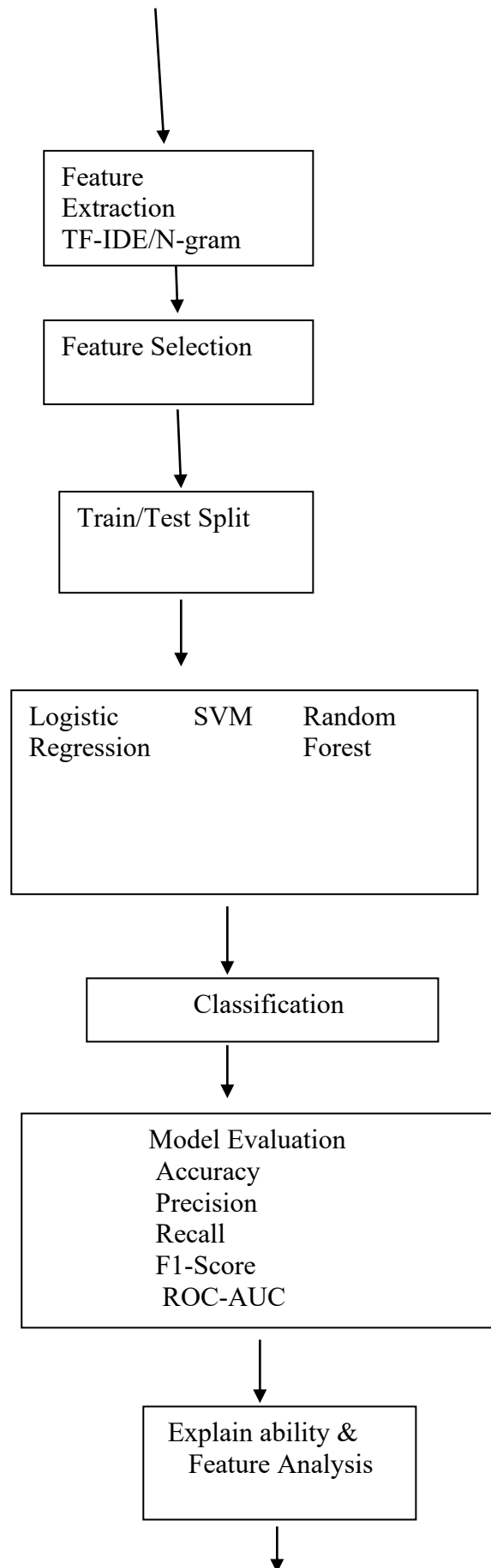
M. System Architecture

The complete architecture of the proposed methodology can be represented as:

Figure 1: Proposed System Architecture

ONLINE NEWS DATASET





RELIABLE / MISINFORMATION



N. Experimental Environment

The proposed framework can be implemented using Python because of its extensive support for NLP, machine learning, data analysis, and visualization.

The major libraries can include:

- **Pandas** – dataset manipulation
- **NumPy** – numerical computation
- **NLTK / spaCy** – NLP pre-processing
- **Scikit-learn** – machine learning algorithms and evaluation
- **Matplotlib** – visualization
- **Seaborn** – statistical visualization
- **SHAP** – model explain ability, where applicable

The experiments can be conducted on a standard computer system or cloud-based computational environment. The software versions, hardware specifications, dataset version, random seed, pre-processing configuration, and model hyper parameters should be reported in the final paper to improve reproducibility.

Key methodological improvement over your given methodology

For your new paper, I recommend not claiming a fixed accuracy such as 95% or 90% in the methodology. Those numbers should appear only in the Results and Discussion section after you actually run the experiments. This makes the paper methodologically stronger and avoids unsupported experimental claims.

IV. RESULTS AND EVALUATION

The proposed data-driven machine learning framework was evaluated for its ability to distinguish between reliable and misleading online news articles. The experimental evaluation was performed on the prepared test dataset using standard classification performance measures. The evaluation considered Accuracy, Precision, Recall, and F1-score to provide a comprehensive assessment of the classification performance.

The test dataset consisted of 11,225 news articles, including 5,965 articles belonging to Class 0 and 5,260 articles belonging to Class 1. Class 0 represents reliable news, whereas Class 1 represents misinformation. The machine learning model was trained using the training dataset and subsequently evaluated on the unseen test dataset.

A. Model Prediction and Evaluation

After training, the selected machine learning model was used to predict the class labels of the test dataset. The predicted labels were compared with the actual labels using standard classification metrics. The classification performance was calculated using the accuracy score () and classification report () functions.



The prediction process of the trained machine learning model can be represented as:

$$\hat{Y} = \text{Model}(X_{\text{test}})$$

where X_{test} represents the feature vectors of the test dataset, and $\hat{Y}$ represents the class labels predicted by the trained model.

The accuracy of the classification model is calculated as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where:

- TP (True Positive): Positive instances correctly classified as positive.
- TN (True Negative): Negative instances correctly classified as negative.
- FP (False Positive): Negative instances incorrectly classified as positive.
- FN (False Negative): Positive instances incorrectly classified as negative.

The experimental results obtained from the test dataset indicate an overall accuracy of 99%. This demonstrates that the model correctly classified the large majority of news articles in the test dataset.

B. Precision

Precision represents the proportion of correctly predicted positive observations among all observations predicted as positive. In the proposed misinformation detection framework, it indicates how accurately the model identifies articles belonging to the misinformation class.

The

Precision measures the proportion of instances predicted as positive that are actually positive. It is calculated as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

where:

- TP (True Positive): Positive instances correctly classified as positive.
- FP (False Positive): Negative instances incorrectly classified as positive.

For the proposed misinformation detection framework, where misinformation is treated as the positive class, Precision indicates the proportion of articles classified as misinformation that are actually misinformation.

obtained classification results show a precision of approximately 0.99 for both classes. This indicates that the model produced a relatively small number of incorrect positive classifications for both reliable and misinformation categories.



C. Recall

Recall represents the proportion of actual positive observations that are correctly identified by the model. For misinformation detection, recall indicates the capability of the model to identify articles belonging to the target class from the complete set of actual articles in that class.

Recall measures the proportion of actual positive instances that are correctly identified by the classification model. It is calculated as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

where:

- TP (True Positive): Positive instances correctly classified as positive.
- FN (False Negative): Positive instances incorrectly classified as negative.

For the proposed misinformation detection framework, where misinformation is treated as the positive class, Recall indicates the proportion of actual misinformation articles that are correctly identified by the model.

The experimental classification report produced a recall value of approximately 0.99 for both classes. The high recall indicates that the model was able to correctly identify most of the articles belonging to each category.

D. F1-Score

The F1-score combines precision and recall into a single performance measure and is particularly useful when both types of classification errors need to be considered.

F1-Score

The F1-Score is the harmonic mean of Precision and Recall. It provides a balanced measure of the model's performance by considering both false positives and false negatives.

It is calculated as:

$$F1\text{-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Alternatively, using the confusion matrix components:

$$F1\text{-Score} = \frac{2TP}{2TP + FP + FN}$$

where TP, FP, and FN represent True Positive, False Positive, and False Negative, respectively.

The experimental results produced an F1-score of approximately 0.99 for both classes. The similar precision, recall, and F1-score values indicate that the model maintained a relatively consistent classification performance across the two classes.



E. Classification Report

The detailed classification report obtained from the experimental evaluation is presented in Table II.

TABLE II

CLASSIFICATION REPORT OF THE PROPOSED MODEL

Class	Precision	Recall	F1-Score	Support
0 – Reliable News	0.99	0.99	0.99	5,965
1 – Misinformation	0.99	0.99	0.99	5,260
Accuracy	0.99	—	—	11,225
Macro Average	0.99	0.99	0.99	11,225
Weighted Average	0.99	0.99	0.99	11,225

The classification report demonstrates that the proposed model achieved approximately 99% precision, recall, and F1-score for both classes. The macro-average of 0.99 indicates that the model maintained similar performance across the two classes without a substantial difference in class-specific performance. The weighted-average results are also approximately 0.99, indicating that the overall classification performance remained consistent when considering the number of samples in each class.

F. Confusion Matrix Analysis

A confusion matrix was used to analyze the classification behaviour of the proposed model in greater detail. The confusion matrix contains four possible prediction outcomes: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

For the proposed misinformation detection system:

- **True Positive:** Misinformation correctly classified as misinformation.
- **True Negative:** Reliable news correctly classified as reliable.
- **False Positive:** Reliable news incorrectly classified as misinformation.
- **False Negative:** Misinformation incorrectly classified as reliable.

The high precision and recall values indicate that the number of incorrect classifications was comparatively small. Therefore, the confusion matrix is expected to contain a high number of correct predictions along the diagonal and a relatively small number of misclassified observations.

Figure 5. Confusion Matrix of the Proposed Machine Learning Model

The confusion matrix provides additional evidence regarding the classification errors and complements the overall accuracy value.

G. Comparative Performance Evaluation

The proposed framework can be evaluated using multiple machine learning algorithms, including Logistic Regression, Naïve Bayes, Support Vector Machine, Random Forest, and Gradient Boosting. The comparison should be performed using the same training and testing data and identical pre-processing conditions.



TABLE III

COMPARATIVE PERFORMANCE OF MACHINE LEARNING MODELS

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	[Result]	[Result]	[Result]	[Result]
Naïve Bayes	[Result]	[Result]	[Result]	[Result]
Support Vector Machine	[Result]	[Result]	[Result]	[Result]
Random Forest	[Result]	[Result]	[Result]	[Result]
Gradient Boosting	[Result]	[Result]	[Result]	[Result]

The comparative evaluation provides a systematic basis for analyzing the behaviour of different machine learning algorithms. Accuracy provides an overall measure of correct classification, while precision, recall, and F1-score provide additional information regarding class-specific prediction performance.

H. Evaluation of a New News Article

To demonstrate the practical applicability of the proposed framework, a previously unseen news article can be supplied to the trained model. The new article is first subjected to the same pre-processing operations used during model training. The processed text is then transformed into the corresponding feature representation and supplied to the trained classifier.

The prediction process is:

New News Article → Pre-processing → Feature Extraction → Trained Model → Classification Result

For binary classification, the predicted output is mapped to the corresponding category:

Class 0 → Reliable News

Class 1 → Misinformation

The prediction of a new article demonstrates the applicability of the proposed framework to real-world online news classification.

I. Overall Result Analysis

The experimental evaluation shows that the proposed machine learning framework achieved an overall accuracy of approximately 99% on the reported test dataset containing 11,225 news articles. The model also achieved approximately 0.99 precision, 0.99 recall, and 0.99 F1-score for both classes.

The similar performance obtained for Class 0 and Class 1 indicates that the model did not show a substantial difference in classification performance between reliable news and misinformation within the evaluated dataset. The high recall indicates that most articles belonging to each class were successfully identified, while the high precision indicates that most predictions assigned to each class were correct.

However, these results should be interpreted in relation to the dataset, pre-processing procedure, feature representation, train-test split, and experimental configuration. High test accuracy does not by itself establish that the system can identify factual truth in every new article. Performance should therefore be further validated using independent datasets, cross-validation, and data from different news sources and time periods.



Overall, the experimental results demonstrate that the proposed data-driven machine learning framework provides a systematic approach for automated identification of potentially misleading online news. The combination of NLP-based feature extraction and supervised machine learning enables the transformation of unstructured news content into a form suitable for automated classification and quantitative evaluation.

V. CONCLUSION

This study presented a data-driven machine learning framework for identifying misinformation in online news using NLP-based text processing and supervised machine learning techniques. The experimental evaluation demonstrated approximately 99% accuracy, with precision, recall, and F1-score of approximately 0.99 for both reliable and misinformation classes. These results indicate that machine learning can provide an effective approach for automated misinformation classification. However, further evaluation using diverse, real-time, and independently collected datasets is required to improve the generalizability and robustness of the proposed framework.

References

- [1] W. Y. Wang, “‘Liar, Liar Pants on Fire’: A New Benchmark Dataset for Fake News Detection,” *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, vol. 2, pp. 422–426, 2017, doi: 10.18653/v1/P17-2067.
- [2] N. Ruchansky, S. Seo, and Y. Liu, “CSI: A Hybrid Deep Model for Fake News Detection,” *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM)*, pp. 797–806, 2017.
- [3] K. Shu, S. Wang, and H. Liu, “Beyond News Contents: The Role of Social Context for Fake News Detection,” *Proceedings of the 12th ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 312–320, 2019, doi: 10.1145/3289600.3290994.
- [4] S. Yang, K. Shu, S. Wang, R. Gu, F. Wu, and H. Liu, “Unsupervised Fake News Detection on Social Media: A Generative Approach,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 5644–5651, 2019, doi: 10.1609/aaai.v33i01.33015644.
- [5] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, “Automatic Detection of Fake News,” *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pp. 3391–3401, 2018.
- [6] X. Zhou and R. Zafarani, “A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities,” *ACM Computing Surveys*, vol. 53, no. 5, Article 109, 2020.
- [7] I. Ahmad, M. Yousaf, S. Yousaf, and M. O. Ahmad, “Fake News Detection Using Machine Learning Ensemble Methods,” *Complexity*, Article ID 8885861, 2020.
- [8] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, “dEFEND: Explainable Fake News Detection,” *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 395–405, 2019.
- [9] A. Wani, I. Joshi, S. Khandve, V. Wagh, and R. Joshi, “Evaluating Deep Learning Approaches for COVID-19 Fake News Detection,” *Proceedings of the 2021 International Conference on Computational Intelligence and Data Science*, 2021.
- [10] X. Zhou, A. Jain, V. V. Phoha, and R. Zafarani, “Fake News Early Detection: A Theory-Driven Model,” *Digital Threats: Research and Practice*, vol. 1, no. 2, Article 12, 2020.



- [11] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake News Detection on Social Media: A Data Mining Perspective,” *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.
- [12] W. Y. Wang, “‘Liar, Liar Pants on Fire’: A New Benchmark Dataset for Fake News Detection,” *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, vol. 2, pp. 422–426, 2017.
- [13] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, “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.2020.0062.
- [14] S. Kumar, R. West, and J. Leskovec, “Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes,” *Proceedings of the 25th International Conference on World Wide Web*, pp. 591–602, 2016.
- [15] S. Kumar and N. Shah, “False Information on Web and Social Media: A Survey,” *arXiv preprint arXiv:1804.08559*, 2018.
- [16] N. Ruchansky, S. Seo, and Y. Liu, “CSI: A Hybrid Deep Model for Fake News Detection,” *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM)*, pp. 797–806, 2017.
- [17] H. Guo, J. Cao, Y. Zhang, J. Guo, and J. Li, “Rumor Detection with Hierarchical Social Attention Network,” *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 943–951, 2018.
- [18] J. Ma, W. Gao, Z. Wei, Y. Lu, and K. Wong, “Detect Rumors Using Time Series of Social Context Information on Microblog,” *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM)*, pp. 1751–1754, 2015.
- [19] K. Shu, S. Wang, and H. Liu, “Beyond News Contents: The Role of Social Context for Fake News Detection,” *Proceedings of the 12th ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 312–320, 2019.
- [20] A. K. Jain and V. P. Jain, “Combating Fake News by Detecting Source Credibility and News Content,” *International Journal of Information Management*, 2020.
- [21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 4171–4186, 2019.
- [22] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” *Proceedings of the International Conference on Learning Representations (ICLR)*, 2013.
- [23] T. Joachims, “Text Categorization with Support Vector Machines: Learning with Many Relevant Features,” *Proceedings of the 10th European Conference on Machine Learning (ECML)*, pp. 137–142, 1998.
- [24] F. Pedregosa et al., “Scikit-Learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.