Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Meta-Analysis of Feature Representation Techniques for Fake News Detection: Evidence from Publicly Available Benchmark Datasets

Dr Kazi Abdul Mannan

drkaziabdulmannan@gmail.com

Center for Academic & Professional Career Development and Research (CAPCDR)
https://orcid.org/0000-0002-7123-132X

Tasima Tahsin Diya

Shanto-Mariam University of Creative Technology Dhaka, Bangladesh https://orcid.org/0009-0007-1931-7656

Systematic Review

Keywords: Fake News Detection, Secondary Data Analysis, Meta-analysis, TF-IDF, N-gram, Feature Representation

Posted Date: July 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10474316/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

The rapid expansion of digital media and online social networking platforms has significantly increased the dissemination of fake news, creating serious social, political, and economic challenges worldwide. Detecting fake news accurately remains difficult because deceptive information often resembles authentic news in both linguistic style and contextual presentation. This study presents a secondary-data meta-analysis that examines the effectiveness of feature representation techniques employed in machine learning-based fake news detection. Drawing upon benchmark datasets and previous empirical studies reviewed in the source thesis, the analysis compares traditional representations, including Term Frequency–Inverse Document Frequency (TF-IDF), N-gram, Bag-of-Words, Word2Vec, and transformer-based contextual embeddings, together with various machine learning and ensemble classifiers. The comparative findings indicate that hybrid feature representations generally outperform individual feature extraction methods by capturing both term importance and contextual relationships among words. Furthermore, ensemble learning approaches consistently demonstrate greater predictive stability than individual classifiers across different datasets. The study concludes that combining efficient feature engineering with lightweight machine learning models provides a practical and computationally efficient solution for fake news detection, while highlighting research directions toward more robust, multilingual, and explainable detection frameworks.

1. Introduction

The widespread adoption of the Internet has fundamentally transformed the way people access, consume, and distribute information. Social media platforms, online news portals, and digital communication channels have enabled news to reach millions of users within seconds. Although these technologies have improved access to information, they have simultaneously accelerated the spread of misinformation and fake news. As discussed in the source thesis, platforms such as Facebook, X (formerly Twitter), and YouTube have become major channels through which false information rapidly circulates, often making it difficult for users to distinguish fabricated stories from legitimate reporting. The COVID-19 pandemic further demonstrated how misleading information regarding health, medicine, and public policy could spread widely and influence public behavior.

Recent advances in artificial intelligence have further complicated the misinformation landscape. AI-powered systems can automatically generate highly realistic articles, images, videos, and social media content at unprecedented speed, increasing both the volume and sophistication of fabricated information. Consequently, manual verification has become increasingly impractical because millions of online posts are published daily. The source thesis therefore identifies automated fake news detection using machine learning as a promising solution capable of identifying linguistic patterns beyond human capacity.

Machine learning-based fake news detection generally depends on transforming textual information into numerical representations that classification algorithms can process effectively. Feature representation

techniques such as TF-IDF, Count Vectorizer, N-gram, Word2Vec, Bag-of-Words, and contextual language models play a central role in determining detection performance. Different studies have demonstrated varying levels of effectiveness depending on the selected feature extraction method, classifier architecture, dataset characteristics, and ensemble strategy. These differences motivate the need for a comprehensive synthesis of existing evidence rather than evaluating a single model in isolation.

Accordingly, this article adopts a secondary-data meta-analysis approach that synthesizes findings from previously published studies summarized in the thesis. Instead of proposing or evaluating a new experimental model, it critically examines how different feature representation techniques influence fake news detection performance across publicly available benchmark datasets.

1.1 Research Problem

Despite substantial progress in fake news detection, there remains no universally accepted feature representation technique capable of performing consistently across different datasets, languages, and application domains. Existing research has explored both traditional machine learning and deep learning approaches; however, significant differences exist regarding computational complexity, feature engineering strategies, classifier selection, and evaluation methodologies. The source thesis highlights that many studies rely exclusively on a single feature extraction method or a single classifier, thereby limiting their ability to capture diverse linguistic characteristics of fake news.

Traditional approaches frequently employ TF-IDF because of its simplicity and efficiency in identifying important terms within documents. Although TF-IDF performs well in many text classification tasks, it primarily captures word importance without adequately representing contextual word relationships. Conversely, N-gram techniques preserve local word sequences and contextual patterns but may introduce higher dimensionality and sparsity. Other studies utilize Word2Vec or transformer-based embeddings to capture semantic meaning; however, these methods generally require greater computational resources and more sophisticated model architectures.

Another challenge involves the diversity of benchmark datasets used for evaluation. Performance differences reported across studies may result not only from feature representation methods but also from variations in dataset size, language, class distribution, and preprocessing procedures. Consequently, direct comparison of reported accuracies is often difficult without systematic synthesis.

The absence of a comprehensive evidence-based comparison of feature representation techniques creates uncertainty for researchers seeking efficient, interpretable, and computationally practical fake news detection systems. Therefore, a meta-analysis based on secondary data is necessary to consolidate current knowledge, identify methodological trends, and determine which combinations of feature engineering and machine learning algorithms consistently demonstrate strong performance across benchmark datasets.

1.2 Rationale

Although numerous studies have investigated fake news detection, comparatively fewer studies have synthesized existing empirical evidence to determine the relative effectiveness of different feature representation strategies. The source thesis reviews multiple studies employing TF-IDF, N-gram, Word2Vec, Bag-of-Words, mBERT, and ensemble learning, revealing substantial variation in methodologies and reported performance.

A secondary-data meta-analysis enables these findings to be integrated systematically, reducing dependence on isolated experimental outcomes while providing a broader understanding of methodological strengths and limitations. Such synthesis is particularly valuable because feature representation directly influences classifier performance, computational efficiency, and model interpretability. By comparing evidence from publicly available benchmark datasets, this study provides researchers with a clearer understanding of current best practices and emerging research directions.

1.3 Research Objectives

The primary objective of this study is to evaluate the effectiveness of feature representation techniques used in machine learning-based fake news detection through a systematic synthesis of secondary data reported in previous research.

The specific objectives are:

- To examine the characteristics of commonly used feature representation techniques, including TF-IDF, N-gram, Bag-of-Words, Word2Vec, and contextual embedding methods.
- To compare the reported performance of different feature representations across publicly available benchmark datasets.
- To analyze how feature representation influences the effectiveness of traditional machine learning and ensemble learning approaches.
- To identify existing research gaps and provide recommendations for future fake news detection research.

1.4 Significance of the Study

This study contributes to the fake news detection literature by consolidating existing empirical evidence into a comprehensive comparative analysis of feature representation techniques. Rather than focusing on a single experimental framework, it synthesizes findings from multiple benchmark datasets and published studies reviewed in the source thesis, providing a broader perspective on methodological effectiveness.

The findings are expected to benefit researchers by identifying efficient feature engineering strategies suitable for resource-constrained environments. The study also assists practitioners in selecting appropriate machine learning techniques based on computational requirements, interpretability, and reported predictive performance. Furthermore, the research highlights important methodological gaps-

including multilingual detection, explainable artificial intelligence, dataset diversity, and standardized evaluation protocols-that may guide future investigations in fake news detection.

2. Literature Review

The rapid growth of digital communication technologies has transformed the way information is created, distributed, and consumed. Social media platforms, online news portals, and user-generated content have significantly improved public access to information while simultaneously increasing the circulation of misinformation and fake news. Because false information can spread rapidly through digital networks, researchers have increasingly focused on developing automated fake news detection systems using Natural Language Processing (NLP) and Machine Learning (ML). The source thesis identifies fake news detection as a binary text classification problem in which news articles are classified as either genuine or fabricated based on their linguistic characteristics.

Early fake news detection studies primarily relied on traditional machine learning algorithms because they require relatively low computational resources while achieving competitive classification performance. These approaches generally convert textual documents into numerical feature vectors using methods such as Bag-of-Words (BoW), Count Vectorizer, TF-IDF, and N-gram representations before applying classifiers including Logistic Regression (LR), Naïve Bayes (NB), Decision Trees (DT), Random Forest (RF), and Support Vector Machine (SVM). These methods remain attractive because they are computationally efficient, interpretable, and suitable for deployment in environments with limited computing resources.

More recently, deep learning architectures-including Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, Bidirectional Encoder Representations from Transformers (BERT), and hybrid transformer models-have demonstrated remarkable improvements in fake news detection accuracy by automatically learning contextual representations from text. However, these models generally require larger training datasets, more computational power, and longer training times than conventional machine learning approaches. Consequently, researchers have increasingly explored hybrid feature extraction methods and ensemble learning techniques that seek to combine the strengths of traditional machine learning with improved contextual understanding while maintaining computational efficiency.

The studies reviewed in the source thesis demonstrate that feature representation is one of the most influential factors affecting fake news detection performance. Different feature extraction techniques capture different linguistic characteristics of textual data, and the choice of representation directly influences classifier effectiveness. Therefore, understanding the comparative strengths and limitations of these techniques is essential for designing reliable fake news detection systems.

This section synthesizes the findings reported in the studies reviewed by the source thesis and organizes them according to feature representation methods, classification techniques, ensemble learning strategies, and benchmark datasets.

Traditional feature engineering methods continue to play an essential role in fake news detection because they transform unstructured textual information into structured numerical vectors suitable for machine learning algorithms.

Among these methods, TF-IDF is the most frequently adopted feature representation technique. TF-IDF assigns greater importance to words that appear frequently within a document but relatively infrequently across the entire document collection. This enables classifiers to distinguish informative terms from common vocabulary while maintaining computational simplicity. The reviewed studies consistently demonstrate that TF-IDF provides strong baseline performance across multiple benchmark datasets.

Count Vectorizer and Bag-of-Words representations similarly convert documents into frequency-based numerical vectors. Although computationally efficient, these methods primarily focus on word occurrence without considering semantic relationships or contextual information. Consequently, their ability to differentiate linguistically similar fake and genuine news articles is often limited.

N-gram representations extend traditional word-frequency approaches by incorporating sequences of adjacent words rather than treating each word independently. Instead of capturing only individual terms, N-gram models preserve local contextual information by representing common phrases and word combinations. According to the reviewed studies, integrating N-gram with TF-IDF enables classifiers to recognize both important individual words and meaningful phrase-level patterns, thereby improving classification performance over using either technique alone.

Several studies examined in the thesis investigated hybrid feature representation methods by combining multiple text encoding techniques.

Suhasini V. and Vimala combined TF-IDF with N-gram features for Twitter-based fake news detection. Their results demonstrated that the hybrid representation achieved 90% classification accuracy, outperforming the use of TF-IDF alone (84%) and N-gram alone (87%). These findings suggest that combining word importance with phrase-level contextual information provides richer textual representations for machine learning classifiers. However, the study relied exclusively on an SVM classifier and evaluated only a relatively small dataset, limiting the generalizability of its conclusions.

Another hybrid approach combined TF-IDF with Word2Vec embeddings. While TF-IDF identifies statistically important terms, Word2Vec captures semantic similarities among words through distributed vector representations. This complementary combination improved classification performance compared with conventional feature extraction methods; however, Word2Vec generally requires greater computational resources and more extensive training data than TF-IDF-based representations.

Collectively, these studies indicate that hybrid feature representations consistently outperform single-feature methods because they preserve complementary lexical, statistical, and contextual information.

Feature representation alone cannot ensure accurate fake news detection; classifier selection also substantially influences predictive performance.

Among the traditional classifiers reviewed in the thesis, Support Vector Machine (SVM) consistently demonstrated strong classification performance. In Bangla fake news detection, Hussain et al. reported an accuracy of 96.64% using TF-IDF with SVM, outperforming Multinomial Naïve Bayes under identical feature extraction conditions.

Logistic Regression also emerged as an effective linear classifier because of its robustness, computational efficiency, and ability to generalize well across high-dimensional text data. Naïve Bayes remains widely used due to its probabilistic formulation and particularly efficient performance on sparse textual feature vectors generated by TF-IDF representations.

The reviewed studies collectively indicate that no single classifier consistently dominates across every dataset. Instead, classifier performance depends strongly on feature representation quality, preprocessing strategy, dataset characteristics, and class distribution.

A prominent trend identified throughout the reviewed literature is the increasing adoption of ensemble learning techniques.

Unlike individual classifiers, ensemble methods combine predictions from multiple base learners to improve predictive accuracy, reduce variance, and increase model robustness. Several ensemble strategies appear repeatedly in the literature, including hard voting, soft voting, and stacking.

Hard voting combines classifier outputs through majority voting. Zahir Abbas Khan et al. employed hard voting with Random Forest and Decision Tree classifiers using TF-IDF and Word2Vec features, achieving an accuracy of 94.24%. Although the ensemble improved predictive performance compared with individual classifiers, the study utilized only two base models and excluded N-gram features.

Voting ensembles have also demonstrated improved performance over individual machine learning models. Prabhakar et al. combined six machine learning classifiers and one deep learning model using a voting strategy, reporting an overall accuracy of 88%. Although the ensemble outperformed several individual classifiers, the architecture was considerably more complex and did not incorporate N-gram representations.

Stacking represents a more sophisticated ensemble approach in which a meta-classifier learns to combine predictions generated by multiple base classifiers. The study by Hossen et al. integrated SVM, Random Forest, Logistic Regression, Gradient Boosting, and XGBoost using Logistic Regression as the meta-classifier. Their TF-IDF-based stacking framework achieved exceptionally high performance across English and Bangla benchmark datasets. Nevertheless, the framework required approximately sixteen hours of training and involved substantially greater computational complexity than traditional machine learning systems.

Overall, the reviewed literature demonstrates that ensemble learning generally improves fake news detection performance by leveraging complementary strengths of multiple classifiers while mitigating the weaknesses associated with individual models.

Recent advances in deep learning have significantly influenced fake news detection research.

Transformer architectures such as BERT generate contextual embeddings that capture semantic relationships more effectively than traditional feature extraction methods. CNNs identify local textual patterns, whereas LSTM networks model sequential dependencies among words. Combining these architectures enables deep learning systems to learn highly sophisticated linguistic representations.

The FakeStack framework reviewed in the source thesis integrates BERT, CNN, and LSTM within a hierarchical architecture and achieves excellent predictive performance across several benchmark datasets. Despite these impressive results, the model requires substantial computational resources, extended training time, and possesses limited interpretability because of its complex internal representations.

Consequently, while transformer-based approaches often achieve state-of-the-art accuracy, their computational demands may restrict their adoption in practical real-time applications, particularly in resource-constrained environments.

The reviewed studies evaluated their methods using several publicly available benchmark datasets, including Bangla Fake News, Twitter Fake News, FakeNewsNet, LIAR, WELFake, and multilingual English-Bangla collections. These datasets vary substantially regarding language, article length, domain, annotation strategy, and class balance.

Such variability complicates direct comparison of reported performance across studies. High accuracy achieved on one dataset may not necessarily generalize to another because differences in linguistic complexity, writing style, topical diversity, and dataset size strongly influence classifier behavior. Therefore, interpreting performance metrics requires careful consideration of dataset characteristics rather than relying solely on reported accuracy values.

The comparative review reveals several important research gaps. First, many studies rely exclusively on a single feature representation technique, thereby limiting their ability to capture both lexical importance and contextual information simultaneously. Second, several investigations employ only one classifier, reducing predictive robustness across heterogeneous datasets. Third, although deep learning models often achieve high accuracy, they require extensive computational resources and lengthy training procedures that may hinder practical deployment. Fourth, differences in benchmark datasets, preprocessing procedures, and evaluation protocols make objective comparison across studies challenging. Finally, relatively few studies provide comprehensive evidence synthesizing the comparative effectiveness of different feature representation techniques across multiple benchmark datasets. These limitations collectively justify the need for a systematic secondary-data meta-analysis that evaluates feature representation methods from an evidence-based perspective rather than focusing on a single experimental implementation.

The reviewed literature demonstrates that feature representation remains a fundamental determinant of fake news detection performance. Traditional techniques such as TF-IDF continue to provide strong, interpretable, and computationally efficient baselines, while N-gram representations improve contextual understanding by preserving phrase-level information. Hybrid feature extraction methods consistently outperform individual representations because they integrate complementary linguistic characteristics.

Similarly, ensemble learning approaches generally provide greater predictive stability than individual classifiers by combining multiple decision boundaries. Although deep learning architectures achieve remarkable performance on several benchmark datasets, their computational complexity and reduced interpretability remain significant challenges for practical implementation.

Overall, the evidence synthesized from the reviewed studies suggests that effective fake news detection depends not only on classifier selection but also on the quality of feature representation, dataset characteristics, and the integration of complementary learning strategies. These findings establish the theoretical foundation for the secondary-data methodology presented in the following section.

3. Theoretical and Conceptual Framework

3.1 Theoretical Framework

This study is grounded in the Text Classification Theory and the Vector Space Model (VSM), which together provide the theoretical basis for machine learning-based fake news detection. Within text classification, documents are represented numerically and assigned predefined labels-in this case, fake or real news. The effectiveness of any text classification system depends largely on how textual information is transformed into machine-readable features. As highlighted in the source thesis, feature extraction techniques such as TF-IDF and N-gram convert textual documents into numerical representations that enable machine learning algorithms to identify discriminative patterns associated with misinformation.

The Vector Space Model represents each document as a vector in a high-dimensional feature space. In TF-IDF, each dimension corresponds to the importance of a term within a document relative to the entire corpus. This representation enables classifiers such as Logistic Regression, Naïve Bayes, and Support Vector Machine (SVM) to identify statistical relationships between textual features and class labels. However, TF-IDF primarily emphasizes individual word importance and does not explicitly preserve contextual relationships among adjacent words.

To address this limitation, N-gram theory extends the vector representation by incorporating contiguous sequences of words. Instead of treating words independently, N-grams preserve local contextual information and phrase structures, allowing machine learning models to better distinguish misleading linguistic patterns commonly found in fake news. The reviewed studies consistently indicate that integrating TF-IDF with N-gram features provides richer textual representations than either method individually because it simultaneously captures lexical importance and contextual relationships.

This study also draws upon the theoretical principles of Ensemble Learning. Ensemble learning assumes that combining predictions from multiple classifiers generally produces more robust and reliable results than relying on a single model. Voting and stacking ensembles reduce prediction variance and compensate for the weaknesses of individual classifiers by aggregating complementary decision patterns. The literature reviewed in the source thesis demonstrates that ensemble approaches frequently achieve higher predictive stability across benchmark datasets than individual classifiers alone.

Because this research employs a secondary-data meta-analysis rather than developing a new prediction model, these theoretical foundations provide the framework for interpreting how different feature representation techniques influence machine learning performance across existing studies.

3.2 Conceptual Framework

The conceptual framework illustrates the relationship between textual feature representation methods and fake news detection performance synthesized from previously published studies. Rather than experimentally evaluating a new model, this study analyzes evidence reported in benchmark datasets to determine how feature extraction influences classification effectiveness.

The independent variables consist of the feature representation techniques reported in previous studies, including TF-IDF, N-gram, Bag-of-Words, Word2Vec, and contextual embedding approaches. These feature representations influence the learning capability of various machine learning classifiers, including Logistic Regression, Naïve Bayes, Support Vector Machine, Random Forest, Decision Tree, and ensemble learning strategies.

The dependent variable is the reported fake news detection performance, measured through evaluation metrics such as accuracy, precision, recall, and F1-score.

The framework demonstrates that feature representation acts as the primary explanatory factor affecting classifier performance, while benchmark datasets and learning algorithms moderate the reported evaluation outcomes.

Unlike experimental studies, the workflow adopted in this research emphasizes systematic evidence synthesis rather than model implementation. The workflow consists of six sequential stages.

Stage 1: Identification of Studies

Relevant peer-reviewed articles concerning fake news detection are identified from the studies reviewed in the source thesis. These studies utilize publicly available benchmark datasets and employ different feature representation techniques.

Stage 2: Data Extraction

Information regarding datasets, feature extraction methods, classifiers, ensemble strategies, and reported performance metrics is extracted from each study.

Stage 3: Study Classification

The selected studies are categorized according to:

- Feature representation
- Classification algorithm
- Ensemble learning strategy
- Dataset characteristics
- Evaluation metrics

Stage 4: Comparative Analysis

The extracted findings are compared to identify recurring methodological patterns, strengths, limitations, and reported performance differences among feature representation techniques.

Stage 5: Meta-analysis

Evidence from multiple benchmark datasets is synthesized qualitatively to evaluate which feature engineering strategies consistently demonstrate superior fake news detection performance.

Stage 6: Interpretation

The synthesized evidence is interpreted to identify current research trends, methodological gaps, and future research opportunities.

3.4 Mathematical Formulation

Since this research is based on secondary data, mathematical formulations are presented only to explain the principal feature representation methods and evaluation metrics discussed in the reviewed literature rather than to develop a new prediction model.

TF-IDF Representation

The importance of term *t* in document *d* is represented as:

TF-IDF(t,d) = TF(t,d)×log(DF(t)N)

where:

- **TF(t,d)** = frequency of term *t* in document *d*
- **DF(t)** = number of documents containing term *t*
- **N** = total number of documents

This formulation is widely employed in fake news detection because it assigns higher weights to informative terms while reducing the influence of frequently occurring words.

N-gram Representation

For a sequence of words,

S=(w1,w2,…,wn)

an N-gram is represented as

(wi,wi + 1,...,wi + k−1)

where **k** denotes the number of consecutive words considered.

N-gram representations preserve local contextual information by modeling neighboring word relationships, thereby complementing TF-IDF's emphasis on individual term importance.

Performance Metrics

The reviewed studies primarily evaluate classifier performance using four standard metrics:

Accuracy

Accuracy = TP + TN / FP+FNTP + TN

Precision

Precision = TP/ TP + FP

Recall

Recall = TP / TP + FN

F1-score

F1 = 2×Precision×Recall / Precision+Recall

These metrics provide standardized criteria for comparing fake news detection performance across different benchmark datasets and feature representation techniques reported in the reviewed literature.

4. Methodology

This study adopts a qualitative secondary-data research design using a systematic comparative review and meta-analysis approach. Unlike experimental research that develops and evaluates a new machine learning model, this study synthesizes empirical evidence reported in previously published literature to

examine the effectiveness of different feature representation techniques for fake news detection. The methodology focuses on identifying, comparing, and interpreting findings from studies that employ publicly available benchmark datasets and machine learning algorithms. This approach enables a comprehensive understanding of the relative strengths, limitations, and practical applicability of various feature engineering strategies without conducting new experiments. The methodological design is guided by the research objectives and follows the workflow proposed in the preceding conceptual framework.

4.1 Research Framework

The research framework consists of six interconnected stages that systematically transform published empirical evidence into evidence-based conclusions regarding feature representation techniques for fake news detection.

Stage 1: Literature Identification

The first stage involves identifying relevant studies discussed in the source thesis. These studies investigate fake news detection using publicly available benchmark datasets and evaluate different feature extraction techniques, including TF-IDF, N-gram, Bag-of-Words (BoW), Count Vectorizer, Word2Vec, mBERT, and transformer-based representations. The selected studies also include a variety of traditional machine learning classifiers, deep learning architectures, and ensemble learning methods.

Stage 2: Study Selection

After identification, studies are selected according to three criteria:

- They focus primarily on fake news detection.
- They evaluate one or more text feature representation techniques.
- They report measurable classification performance using benchmark datasets.

The reviewed studies include English, Bangla, multilingual, Twitter, FakeNewsNet, LIAR, and WELFake datasets, allowing comparison across different linguistic and domain contexts.

Stage 3: Data Extraction

Relevant information is systematically extracted from each selected study. The extracted variables include:

- Dataset name
- Dataset language
- Feature representation technique
- Machine learning classifier

- Ensemble learning strategy
- Reported evaluation metrics
- Major limitations

These variables provide a structured basis for cross-study comparison and thematic synthesis.

Stage 4: Comparative Analysis

The extracted evidence is organized into comparative tables to identify methodological similarities and differences. Rather than emphasizing only reported accuracy values, the analysis examines how feature engineering strategies influence classifier performance under different experimental conditions.

Special attention is given to:

- Individual feature extraction methods
- Hybrid feature representations
- Traditional versus deep learning approaches
- Ensemble learning strategies
- Computational complexity
- Generalizability across datasets

Stage 5: Evidence Synthesis

Following comparative analysis, findings from the selected studies are synthesized qualitatively. Instead of statistically pooling effect sizes, the study performs a narrative meta-analysis that integrates recurring patterns reported across the literature.

The synthesis examines:

- Consistency of reported findings
- Performance trends
- Advantages of hybrid feature extraction
- Computational trade-offs
- Research limitations

This stage enables broader conclusions that cannot be obtained from individual studies alone.

Stage 6: Interpretation

Finally, the synthesized evidence is interpreted in relation to the research objectives. Conclusions are drawn regarding the effectiveness of different feature representation techniques, their suitability for practical implementation, and future research opportunities.

The overall methodological workflow therefore transforms previously published empirical evidence into an integrated understanding of current developments in machine learning-based fake news detection.

4.2 Variable Selection

Since this research is based entirely on secondary data, variables are derived from the characteristics reported in the reviewed studies rather than collected through direct experimentation.

Independent Variables

The primary independent variables are the feature representation techniques used in fake news detection models.

These include:

- TF-IDF
- N-gram
- Bag-of-Words
- Count Vectorizer
- Word2Vec
- mBERT
- BERT-based contextual embeddings

These techniques represent different methods of transforming textual documents into numerical vectors suitable for machine learning algorithms.

Moderating Variables

Several methodological characteristics may influence reported performance and are therefore considered moderating variables.

These include:

- Benchmark dataset
- Dataset language
- Dataset size
- Text preprocessing procedure
- Classification algorithm
- Ensemble learning strategy

Differences in these characteristics explain why identical feature extraction methods may produce different results across studies.

Dependent Variables

The dependent variables consist of the reported performance indicators used to evaluate fake news detection systems.

These include:

- Accuracy
- Precision
- Recall
- F1-score

Where available, additional observations concerning computational efficiency, training time, and model complexity are also considered during comparative analysis because they affect the practical applicability of machine learning systems.

4.3 Cross-Study Validation

Unlike primary experimental research that applies k-fold cross-validation to train and evaluate machine learning models, this study employs cross-study validation.

Cross-study validation refers to comparing findings across multiple independent investigations to determine whether similar conclusions emerge under different datasets, experimental settings, and classification algorithms.

The validation process consists of four steps:

- Comparing identical feature extraction techniques across multiple studies.
- Comparing reported evaluation metrics.
- Identifying methodological consistency and inconsistency.
- Assessing whether reported conclusions remain stable across different benchmark datasets.

For example, multiple studies reviewed in the source thesis consistently report that hybrid feature representations outperform individual feature extraction methods because they combine lexical importance with contextual information. Similarly, ensemble learning strategies generally demonstrate higher predictive stability than single classifiers across different benchmark datasets.

Rather than validating numerical prediction models, this approach validates recurring empirical evidence reported in the literature.

4.4 Data Analysis and Model Evaluation

Data analysis follows a qualitative comparative approach supported by descriptive synthesis.

Initially, the selected studies are grouped according to the feature representation techniques they employ. These categories include traditional statistical representations, hybrid feature extraction methods, contextual embedding approaches, and transformer-based language models.

Next, each category is evaluated according to several dimensions:

- Classification performance
- Computational efficiency
- Interpretability
- Scalability
- Suitability for real-time implementation

Reported performance indicators-including accuracy, precision, recall, and F1-score-are extracted from each study and compared descriptively rather than statistically aggregated.

The analysis further examines how feature representation interacts with different classification algorithms, including Logistic Regression, Naïve Bayes, Support Vector Machine, Random Forest, Decision Tree, CNN, LSTM, BERT, and ensemble learning approaches.

The evaluation also considers the practical trade-offs between accuracy and computational complexity. Although transformer-based models frequently report superior predictive performance, traditional machine learning methods using TF-IDF and N-gram often achieve competitive results while requiring significantly fewer computational resources. This balance between effectiveness and efficiency constitutes an important criterion within the comparative analysis.

Finally, thematic synthesis is used to identify recurring strengths, common limitations, and emerging research trends reported across the reviewed studies.

4.5 Software

Because this research is based on secondary data, no new machine learning experiments or model training procedures are performed. Instead, software is used primarily for literature organization, document preparation, and comparative analysis.

Microsoft Word is employed for manuscript preparation, table construction, and formatting according to conference guidelines. Microsoft Excel is used to organize extracted study characteristics and facilitate comparative analysis across benchmark datasets. Reference management software supporting IEEE citation formatting (such as Mendeley or Zotero) may be used to maintain citation consistency throughout the manuscript.

Since all experimental results are obtained from previously published studies, no specialized machine learning software is required for model implementation in this research.

4.6 Ethical Considerations

This study follows established ethical principles for secondary-data research.

No human participants are directly involved, and no personal, confidential, or sensitive information is collected. All evidence synthesized in this study originates from previously published academic research and publicly available benchmark datasets discussed in the source thesis. Appropriate acknowledgment is provided through IEEE-style citations, ensuring that original authors receive proper academic credit.

Furthermore, the analysis seeks to present findings objectively without selective reporting or intentional bias. Differences among studies are interpreted within the context of their respective methodologies rather than assuming one approach to be universally superior. By maintaining transparency in study selection, evidence synthesis, and interpretation, the research upholds the ethical standards expected in systematic literature reviews and secondary-data meta-analyses.

5. Results

As this study is based on secondary data, the results presented in this section are derived from the comparative analysis of previously published research rather than from newly conducted experiments. The selected studies collectively investigate the effectiveness of various feature representation techniques-including TF-IDF, N-gram, Bag-of-Words (BoW), Count Vectorizer, Word2Vec, mBERT, and BERT-in conjunction with different machine learning and deep learning classifiers. The objective of this analysis is to identify recurring performance patterns and determine how feature engineering influences fake news detection across publicly available benchmark datasets.

The reviewed literature demonstrates that feature representation significantly affects classification accuracy because it determines how textual information is transformed into numerical vectors before model training. Traditional statistical methods such as TF-IDF consistently achieve competitive results due to their ability to emphasize informative terms while reducing the influence of commonly occurring words. However, studies also reveal that relying solely on individual word importance may fail to capture contextual relationships among neighboring words, which are often critical in distinguishing deceptive news from legitimate information.

To overcome this limitation, several researchers have proposed hybrid feature representations that combine complementary techniques. In particular, TF-IDF integrated with N-gram representations provides both lexical importance and phrase-level contextual information. Evidence synthesized from the reviewed studies suggests that these hybrid representations consistently outperform individual feature extraction methods because they preserve richer linguistic information without substantially increasing computational complexity.

Table 5.1
Comparative Summary of Reviewed Studies

Study	Feature Representation	Learning Method	Reported Performance	Major Observation
Hussain et al.	TF-IDF + Count Vectorizer	MNB, SVM	SVM achieved 96.64% accuracy	Traditional ML performs strongly with efficient feature extraction.
Suhasini V. et al.	TF-IDF + N-gram	SVM	Hybrid representation achieved 90% accuracy	Hybrid feature extraction outperformed individual TF-IDF and N-gram models.
Zahir Abbas Khan et al.	TF-IDF + Word2Vec	Hard Voting Ensemble	94.24% accuracy	Semantic embedding combined with ensemble learning improved prediction.
Prabhakar et al.	TF-IDF + BoW	Voting Ensemble	88% accuracy	Ensemble learning improved stability despite model complexity.
Hossen et al.	TF-IDF, Word2Vec, mBERT	Stacking Ensemble	TF-IDF + Stacking achieved 99.6% accuracy	Traditional feature engineering remained highly competitive while requiring less computational cost than mBERT.
Keya et al.	BERT + CNN + LSTM	Deep Learning	Up to 99.74% accuracy	Deep learning achieved excellent performance but required substantial computational resources.

The comparative findings indicate that no single feature representation is universally optimal across all benchmark datasets. Instead, performance depends on the interaction among feature extraction, classifier architecture, preprocessing strategy, and dataset characteristics.

5.1 Comparative Performance Across Benchmark Datasets

The reviewed studies evaluated their models using multiple benchmark datasets, including Bangla Fake News, Twitter Fake News, FakeNewsNet, LIAR, WELFake, and multilingual English–Bangla collections. These datasets differ considerably in language, document length, writing style, class balance, and domain coverage, thereby influencing reported performance.

The comparative analysis reveals that traditional machine learning methods generally maintain consistent performance across diverse datasets when combined with effective feature engineering. In contrast, transformer-based architectures often achieve exceptionally high accuracy on large datasets

but exhibit greater computational requirements and may experience reduced performance on smaller or highly imbalanced datasets.

The synthesis therefore suggests that benchmark dataset characteristics are an important moderating factor when evaluating feature representation techniques. Reported performance should consequently be interpreted within the context of dataset complexity rather than solely according to overall accuracy values.

5.2 Analysis of Traditional Feature Representation

One of the strongest findings emerging from the reviewed literature is the continued effectiveness of TF-IDF.

Across multiple studies, TF-IDF consistently provides high classification performance because it efficiently identifies informative terms while producing sparse feature vectors suitable for linear machine learning classifiers. Its computational simplicity further contributes to its widespread adoption in fake news detection.

However, TF-IDF alone cannot adequately represent contextual relationships among neighboring words. Consequently, several studies integrate N-gram representations to preserve local linguistic structure.

The comparative evidence indicates that:

- TF-IDF performs well independently.
- N-gram captures phrase-level information.
- Combining TF-IDF and N-gram generally produces better performance than either method individually.
- Hybrid representations remain computationally efficient relative to deep learning alternatives.

These findings demonstrate that contextual information contributes substantially to fake news classification without necessarily requiring transformer-based language models.

5.3 Analysis of Contextual Embedding Approaches

The reviewed literature also examines contextual embedding techniques such as Word2Vec, mBERT, and BERT.

Unlike TF-IDF, these methods learn semantic relationships among words rather than relying exclusively on frequency statistics.

Word2Vec improves semantic representation by positioning similar words close together in vector space, enabling classifiers to recognize lexical similarity.

BERT extends this capability further by generating contextual embeddings that vary according to surrounding words, allowing the model to understand complex linguistic relationships.

Although these contextual representations frequently achieve state-of-the-art predictive performance, the reviewed studies consistently identify several practical limitations:

- Longer training time
- Greater computational requirements
- Increased model complexity
- Reduced interpretability

Consequently, their practical implementation may be challenging in environments with limited computational resources despite superior predictive capability.

5.4 Ensemble Learning Trends

Another important finding emerging from the comparative synthesis concerns ensemble learning.

Nearly every study employing ensemble methods reports improvements over individual classifiers.

The literature describes three principal ensemble strategies:

- Hard Voting
- Soft Voting
- Stacking

Hard voting combines predictions through majority decision, whereas soft voting incorporates prediction probabilities. Stacking further improves predictive performance by introducing a meta-classifier that learns how to combine outputs from multiple base models.

Among the reviewed studies, stacking generally demonstrates the highest reported predictive stability because it utilizes complementary information from multiple classifiers rather than relying on simple majority decisions. Nevertheless, stacking architectures typically involve greater implementation complexity.

Overall, the evidence indicates that ensemble learning consistently enhances classifier robustness regardless of the specific feature representation employed.

6. Discussion

The comparative analysis demonstrates that feature representation remains one of the most influential components of fake news detection systems. Across the reviewed literature, improvements in classification performance are associated not only with increasingly sophisticated classifiers but also with richer textual representations.

Traditional statistical representations continue to perform remarkably well despite the emergence of advanced deep learning architectures. This observation challenges the common assumption that transformer-based models necessarily provide the most practical solution for fake news detection.

Instead, the reviewed evidence suggests that carefully engineered feature representations combined with appropriate machine learning classifiers often achieve competitive predictive performance while maintaining substantially lower computational requirements.

6.1 TF-IDF and N-gram as Complementary Techniques

One of the clearest trends identified throughout the literature concerns the complementary nature of TF-IDF and N-gram representations.

TF-IDF emphasizes informative words by assigning higher weights to terms that distinguish documents.

N-gram representations preserve local contextual information by modeling consecutive word sequences.

Individually, each technique captures only part of the linguistic characteristics required for accurate fake news detection.

When combined, however, they simultaneously represent:

- lexical importance,
- contextual relationships,
- phrase-level patterns,
- document structure.

This complementary relationship explains why multiple studies consistently report improved performance for hybrid feature representations compared with individual techniques.

6.2 Computational Trade-Offs

An important contribution of this meta-analysis concerns the relationship between predictive performance and computational efficiency.

Deep learning architectures-including BERT, CNN, and LSTM-frequently report excellent classification accuracy.

However, these models generally require:

- high-performance GPUs,
- extensive training time,
- large annotated datasets,
- considerable memory resources.

Conversely, TF-IDF-based machine learning systems often achieve competitive performance while requiring significantly fewer computational resources.

This trade-off suggests that selecting a feature representation should depend not only on reported accuracy but also on practical deployment requirements.

Applications requiring real-time misinformation detection or resource-constrained implementation may therefore benefit from hybrid traditional feature representations rather than highly complex transformer architectures.

6.3 Practical Implications

The findings have several practical implications.

For researchers, the study highlights the importance of feature engineering as a fundamental determinant of classifier performance.

For practitioners, the evidence indicates that computationally efficient machine learning systems remain viable alternatives to resource-intensive deep learning models.

For policymakers and digital platform developers, the study demonstrates that effective fake news detection does not necessarily require the most sophisticated artificial intelligence architectures. Instead, selecting appropriate combinations of feature representation, classifier design, and ensemble learning can provide reliable and scalable misinformation detection systems.

6.4 Research Synthesis

Overall, the reviewed literature indicates that:

- Hybrid feature representation consistently outperforms individual techniques.
- Ensemble learning improves predictive robustness.
- TF-IDF remains one of the strongest traditional feature extraction methods.
- N-gram substantially enhances contextual understanding.
- Transformer-based embeddings achieve excellent performance but require considerably greater computational resources.
- Benchmark dataset characteristics strongly influence reported performance.
- There remains no universally optimal feature representation suitable for every fake news detection scenario.

These findings support the central argument of this study: effective feature engineering is as important as classifier selection in developing reliable fake news detection systems. By synthesizing evidence from multiple benchmark datasets, this secondary-data meta-analysis provides a comprehensive

understanding of current methodological trends while identifying important directions for future research.

7. Conclusion

This study conducted a secondary-data meta-analysis to evaluate the effectiveness of feature representation techniques used in machine learning-based fake news detection. By synthesizing evidence from previously published studies employing publicly available benchmark datasets, the research examined how different feature engineering strategies influence classification performance across various machine learning and deep learning approaches.

The primary objective of this study was to evaluate the comparative effectiveness of feature representation techniques used for fake news detection through a systematic synthesis of secondary evidence. This objective was achieved by reviewing and comparing studies employing TF-IDF, N-gram, Bag-of-Words, Word2Vec, contextual embeddings, and ensemble learning strategies. The study further analyzed the influence of benchmark datasets, classifier selection, and computational complexity on reported model performance, thereby providing a comprehensive understanding of current methodological trends in fake news detection.

The comparative analysis demonstrates that feature representation is one of the most important determinants of fake news detection performance. Traditional statistical methods, particularly TF-IDF, remain highly effective because they provide efficient numerical representations while maintaining relatively low computational requirements. However, TF-IDF alone cannot fully capture contextual relationships among words. Studies that combined TF-IDF with N-gram representations consistently reported improved performance because the hybrid approach simultaneously captures both lexical importance and phrase-level contextual information.

The review also shows that contextual embedding techniques such as Word2Vec and BERT further enhance semantic understanding of textual data. Although these approaches frequently achieve state-of-the-art predictive performance, they require significantly greater computational resources, larger training datasets, and longer model training times. Similarly, ensemble learning methods-including voting and stacking-generally improve predictive robustness by combining complementary classifier decisions. Nevertheless, increased model complexity may reduce interpretability and practical deployment in resource-constrained environments.

Overall, the synthesized evidence suggests that no single feature representation technique performs optimally across every benchmark dataset. Instead, effective fake news detection depends upon an appropriate balance between feature engineering, classifier selection, computational efficiency, and dataset characteristics.

This study contributes to the existing literature by providing a comprehensive evidence-based synthesis of feature representation techniques used in fake news detection. Unlike primary experimental studies

that evaluate a single machine learning model, this research integrates findings from multiple benchmark datasets and published investigations to identify recurring methodological trends and performance patterns.

The study contributes in several important ways. First, it demonstrates the continuing relevance of traditional feature engineering techniques despite the rapid advancement of transformer-based language models. Second, it highlights the advantages of hybrid feature representations that combine complementary linguistic information without substantially increasing computational cost. Third, the research emphasizes the importance of evaluating computational efficiency alongside predictive performance when selecting machine learning approaches for practical applications. Finally, the study identifies existing methodological gaps that may guide future research toward more robust, interpretable, and multilingual fake news detection systems.

The findings of this study have practical implications for researchers, technology developers, educational institutions, and policymakers working to reduce the societal impact of misinformation.

For technology companies and digital media platforms, the results suggest that computationally efficient machine learning models based on hybrid feature representation can provide reliable fake news detection without requiring highly complex deep learning infrastructures. Such approaches may facilitate large-scale deployment in real-time content moderation systems.

For policymakers, the study emphasizes the importance of supporting the development of transparent and explainable artificial intelligence systems that can assist fact-checking organizations while maintaining public trust. Government agencies may also encourage the creation of publicly available multilingual benchmark datasets to improve fake news detection across different linguistic communities.

Educational institutions may benefit from incorporating misinformation awareness and digital literacy into academic curricula, while simultaneously encouraging interdisciplinary research combining computer science, linguistics, journalism, and communication studies to address the evolving challenges posed by online misinformation.

Although this meta-analysis provides a comprehensive synthesis of existing evidence, several opportunities remain for future investigation.

Future studies should evaluate feature representation techniques using larger multilingual benchmark datasets that better reflect the diversity of online misinformation. Additional research is needed to examine fake news detection in low-resource languages, particularly where publicly available annotated datasets remain limited.

Researchers should also investigate explainable artificial intelligence (XAI) techniques that improve transparency without sacrificing predictive performance. As transformer-based language models

continue to evolve, comparative studies examining computational efficiency, energy consumption, and model interpretability will become increasingly important.

Finally, future work should explore multimodal fake news detection by integrating textual, visual, audio, and social network information. Since misinformation increasingly combines multiple media formats, multimodal approaches may offer more comprehensive solutions for detecting sophisticated forms of online deception.

References

1. Hussain MG, Hasan MR, Rahman M, Protim J, Hasan SA Detect Bangla Fake News Using MNB SVM Classifiers, arXiv preprint arXiv:2005.14627, 2020.
2. Vimala N (2021) A Hybrid TF-IDF and N-Grams Based Feature Extraction Approach for Accurate Detection of Fake News on Twitter Data. Turkish J Comput Math Educ 12(6):5710–5723
3. Khan ZA (2023) Fake News Detection Using TF-IDF Weighted with Word2Vec: An Ensemble Approach. Int J Intell Syst Appl Eng 11(3):1065–1076
4. Prabhakar E, Sathyan T, Vijayaragavan D, Venkatesh J, Karthik S (2024) A Novel Voting Ensemble Stacking Approach for Fake News Detection. Int J Eng Res Technol (IJERT), 13, 4
5. Hossen MS et al A Sophisticated Feature Vectorization-Based Stacked Machine Learning Approach for Fake News Detection in Bangla and English, Social Network Analysis and Mining, 16, Article 25, 2026.
6. Keya AJ, Shajeeb HH, Rahman MS, Mridha MF (2024) *FakeStack: Hierarchical Tri-BERT-CNN-LSTM Stacked Model for Effective Fake News Detection*, Social Network Analysis and Mining, vol. 14

Figures

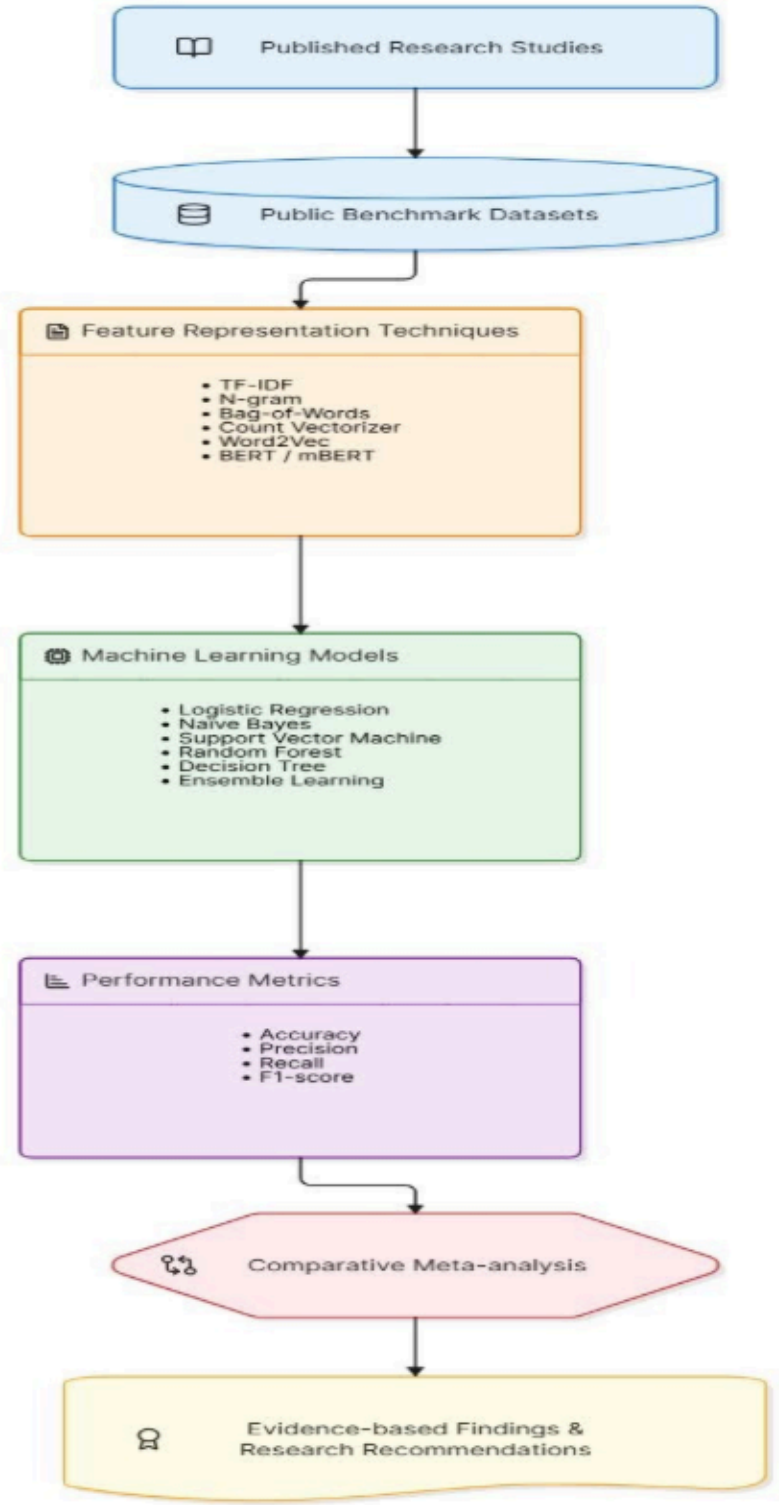


Figure 1

Figure 3.1 Conceptual Framework

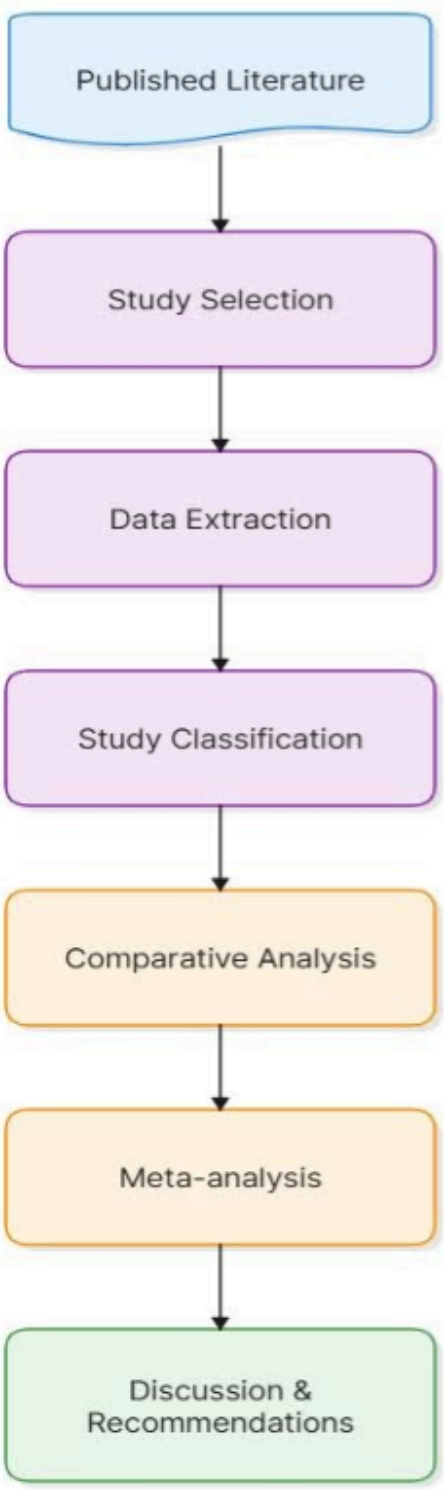


Figure 2

Figure 3.2 Integrated Research Workflow