International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(14s) 2026 pp. 1-22
https://doi.org/10.70917/ijcisim-2026-4193
Article

A Hybrid Semantic–Sentiment Framework for Automatic Fake News Detection Using Doc2Vec and Machine Learning

Sneha Patle[1], Bhushan Gedam[2], Sameer Tembhurney[3], Sachin Balvir[4], Ashwini Agashe[5], Megha Rode[6], Yuvaraj Dakhane[7], Hrushikesh Madhukar Panchabudhe[8]

[1] Department of Computer Science and Engineering, Ramdeobaba University, Nagpur, Maharashtra, India.
Email: patlesk_1@rknec.edu
[2] Department of Computer Science and Engineering, Ramdeobaba University, Nagpur, Maharashtra, India.
Email: bhushangedam44@gmail.com
[3] Department of Computer Science and Engineering, Ramdeobaba University, Nagpur, Maharashtra, India.
Email: sameer.tembhurney@gmail.com
[4] Department of Computer Science and Business Systems, St. Vincent Pallotti College of Engineering and Technology, Nagpur, Maharashtra, India.
Email: sachin_balvir@yahoo.com
[5] Department of Computer Science and Engineering, Ramdeobaba University, Nagpur, Maharashtra, India.
Email: agashear@rknec.edu
[6] Department of Computer Science and Engineering, Ramdeobaba University, Nagpur, Maharashtra, India.
Email: megha.rode14@gmail.com
[7] Senior Automation Engineer, L&T Technology Services, India.
Email: dakhaneyuvaraj@gmail.com
[8] Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International (Deemed University), Pune, Maharashtra, India.
Email: hrushikesh.panchabudhe@sitnagpur.siu.edu.in

Abstract: The spread of misinformation through digital plat-forms such as social media sites and news portals has posed a problem of maintaining information credibility and building trust. Manual approaches alone cannot help cope with the huge volumes of information uploaded on these platforms every day. An automatic approach for fake news detection which utilizes NLP, sentiment analysis, semantic embedding methods and several machine learning algorithms has been developed in this paper. The news headlines collected from FakeNewsNet dataset have been pre-processed via tokenization, stop-word elimination, lemmatization, and n-grams extraction. Doc2Vec approach has been employed to extract semantic vectors whereas sentiment analysis has been done with the help of VADER tool. These semantic vectors have been provided as input to various machine learning algorithms such as Logistic Regression, Linear SVM, Random Forest, Gradient Boosting, XGBoost, LightGBM, Naïve Bayes and K-Nearest Neighbor. Experimental results suggest that ensemble learning models outperform other forms of machine learning techniques. Out of all the tested algorithms, ExtraTrees performed with the highest classification accuracy (77.54%) whereas XGBoost produced the highest macro F1-Score (0.5059)..

Keywords: Fake News Detection, Natural Language Pro-cessing, Doc2Vec, Sentiment Analysis, Machine Learning, XG-Boost, Random Forest, FakeNewsNet..

1. Introduction

Social media website creation has revolutionized the pro-duction, dissemination, and consumption of information. There are millions of people accessing information from social media websites such as Facebook, Twitter, Instagram, YouTube, and even online news websites. Although these platforms provide for easy access to information,

they have become a primary source of fake news and miscommunication. The spread of such false information may affect people's perception and functioning of democracy negatively [1]–[3].

Misleading news, which is referred to as fake news, can be described as a situation where the information provided is passed off as valid news. This is made easy by the way the current communication networks operate since the dissemination is fast and broad. As such, it makes the detection of misleading news a critical issue that should be handled scientifically. However, human-based fact-checkers are costly, slow, and difficult to scale [1], [4], [5].

Development of NLP and ML allowed the creation of au-tomated solutions for detection of deceptive text patterns. Se-mantic representation technologies like Word2Vec, Doc2Vec, GloVe, and transformer-based embeddings proved to be quite effective in obtaining contextual information from textual data. Sentiment analysis allows obtaining additional linguistic insights that may help detect fake news [6]–[9].

The major contributions of this work are summarized as follows:

Development of a complete NLP-based fake news detec-tion pipeline using textual news headlines.

Integration of Doc2Vec semantic embeddings for captur-ing contextual information.

Incorporation of VADER sentiment features to enhance textual characterization.

Comparative evaluation of eight machine learning classi-fiers including ensemble learning approaches.

Comprehensive performance analysis using Accuracy, Precision, Recall, F1-score, and ROC-based evaluation metrics.

Despite the promising performance shown by the proposed approach, some constraints still remain. In the present study, we make use of headlines instead of articles and do not take into account such factors as the impact of multimedia information, user actions, and social dissemination [10]–[12]. The rest of this paper is structured as follows. In Section II, the literature review is provided. In Section III, we discuss the dataset and system models. In Section IV, we describe the proposed methodology. Section V contains experimental results and performance analysis. Finally, Section VI provides

conclusions and outlines future research directions.

2. Literature Review

Considering that disinformation spreads fast over the dis-tributed networks of social media, automation to detect such fake news has become a key necessity for NLP and ML. The existing methods employed in such endeavors to detect fake news can be classified under linguistic engineering, semantic vector representation, deep sequential modeling, and network propagation methods [1], [2], [10].

2.1 Traditional Machine Learning and Linguistic Feature En-gineering

As for the earlier papers on digital fact-checking, a lot of attention was paid to classical machine learning techniques for statistics along with shallow lexical features. The authors employed a sparsely represented language model based on BoW and TF-IDF techniques and classical classifiers like Naïve Bayes, Logistic Regression, and SVM [6], [13], [14].

Despite the fact that these models are not complex in terms of computation power, they require thorough fine-tuning through manual feature engineering that includes readability measures calculation, punctuation rate counting, and POS tagging. These models are very sensitive to vocabulary evasion

attacks because even small changes in the writing style of modern clickbait headlines lead to huge accuracy drop [1], [7].

2.2 Semantic Vector Space Representations and Document Em-beddings

To find out semantic dependencies and structures without manual feature extraction, scientists began to work with dense vector representations. Word2Vec and GloVe algorithms al-lowed obtaining semantic vectors for tokens depending on co-occurrence vectors on the local level [6], [8].

Nevertheless, when working with short texts, for example, with news headlines that involve averaged word vectors, there is a risk of losing context at the document level. To deal with this problem, researchers introduced paragraph embedding algorithms. Among others, the most popular algorithm is Paragraph Vector algorithm of Doc2Vec in distributed memory mode (PV-DM). The Doc2Vec algorithm introduces arbitrary identifiers of documents in the prediction context layer [3], [13].

2.3 Deep Learning and Transformer Architecture Trends

The latest advancements in NLP techniques have been focused on using deep sequence neural models. In particular, the architecture of the convolutional neural network (CNN) model is altered in order to detect the presence of localized spatial n-gram structure in the input text, while RNN and BiLSTM (bidirectional LSTM) networks are widely used to detect temporal structures [7], [8].

However, the recent advances in deep learning have led to development of attention-based mechanisms and transformer models like BERT, RoBERTa, and DistilBERT, which have provided the latest baselines in NLP tasks on large text data. Nevertheless, these models possess huge computational costs and are black box systems [9], [15]–[17].

2.4 Graph Neural Networks and Social Propagation Dynamics

Another type of current research studies the non-textual social context by considering the propagation of fake news as a graph. Graph Neural Network (GNN), Graph Convolutional Network (GCN), and heterogeneous network embedding algo-rithms include using user interaction graph, retweet graph, and publisher profiles to detect any anomalies within the process of spreading the news [10]–[12].

However, social network analysis can detect anomalies in the processes only when there is an active virus thread. Cold start is still a problem because social network analysis cannot detect a deceitful headline until it is spread on the internet. Moreover, a complete collection of user interaction data cannot be collected due to privacy concerns with API [1], [10].

2.5 Linguistic Emotion Profiles and Sentiment Analysis

From the studies made in the field of psycholinguistics, one may conclude that the main purpose of designing deceptive in-formation is to evoke certain emotions from the target audience like fear, anger, or indignation. Hence, linguistic sentiment

distribution measurement becomes a crucial supplement to the feature set [7], [9].

The sentiment analysis using lexical analyzers (Valence Aware Dictionary and sEntiment Reasoner - VADER) gives a great tool for measuring the sentiment intensity in short pieces of text on the social media. The combination of structural and semantic representations of documents with the explicit emotion valence allows taking into consideration the context and the intent of the text writer [6], [7], [9].

3. Literature Summary and Research Gap Matrix

Though a lot of attention from the academic community has been directed towards addressing this issue, a significant gap in research still persists since most of the proposed approaches are geared towards finding deeper semantic relationships but disregard explicit emotional tagging [2], [7], [8].

Table I provides a structured summary matrix tracking core methodologies, dataset baselines, and performance insights across foundational and contemporary literature.

3.1 Research Gap Analysis

Notwithstanding significant advancement in fake news de-tection, there are certain limitations found within the current literature [2], [8], [10]:

Deep learning techniques have high levels of accuracy but need considerable computing power [8], [15], [16].

Graph-based techniques heavily rely on social network data that may not always be accessible [1], [12].

Fact verification methods require access to reliable exter-nal knowledge sources [17], [23].

Several studies focus primarily on either semantic infor-mation or sentiment information rather than integrating both [6], [7], [9].

The interpretability and scalability of advanced fake news detection systems remain challenging [2], [17], [24].In an attempt to overcome these drawbacks, this study suggests a computationally efficient system which integrates Doc2Vec semantic embeddings, VADER sentiment analysis, and a number of classifiers for machine learning. The aim here is to find a compromise between classification accuracy, interpretability, and ease of implementation, using news head-lines only.

Dataset and System Model

This section describes the dataset utilized in this study, the overall system architecture, modeling assumptions, and the notation employed throughout the proposed fake news detection framework.

3.2 Dataset Description

The FakeNewsNet dataset is used for the purpose of exper-imentation because of the extensive usage of the dataset in fake news detection [1], [7], [8]. This data set contains news stories that are taken from different news websites as well as fact-checking websites. Each news story has a label associated

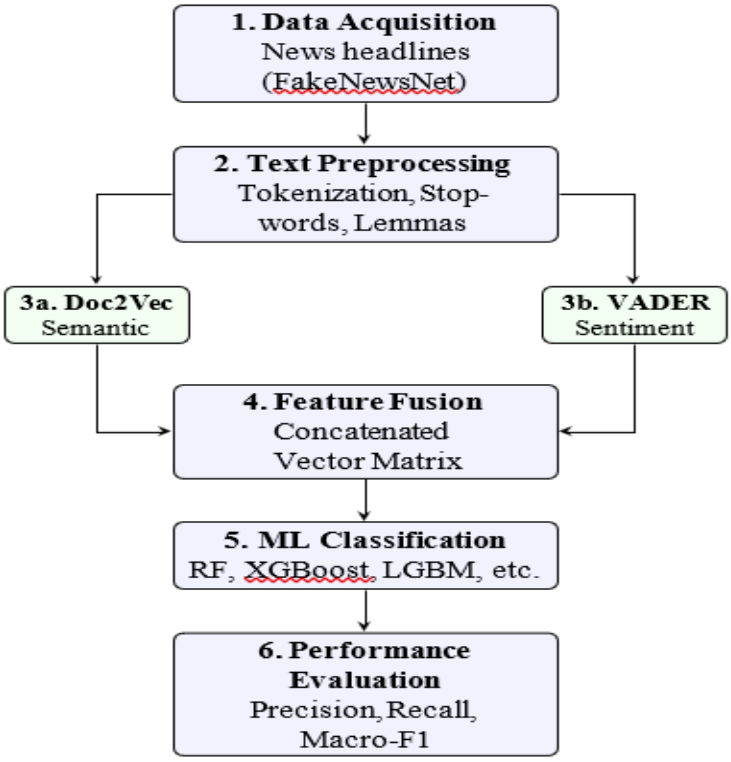

Fig. 1. Overall architecture and processing flow of the proposed fake news detection framework.

with it, which can be one of two possible labels: *Real* or *Fake* [1].There are multiple features in the data set, which include news headline, news URL, details of the publishing company, social engagement details, and the class label. In this study, news headlines will be used for classification purposes [1], [13].

Table III summarizes the dataset characteristics.

The dataset exhibits a class imbalance where real news instances significantly outnumber fake news samples. Con-sequently, evaluation metrics beyond simple accuracy are required to obtain a comprehensive assessment of classifier performance [7], [8], [25].

3.3 System Architecture

The proposed framework consists of five major stages:

Data Acquisition

Text Preprocessing

Feature Extraction

Machine Learning Classification

Performance Evaluation

Figure 1 illustrates the overall workflow of the proposed fake news detection framework.

The processing flow can be summarized as follows:

News headlines are extracted from the FakeNewsNet dataset

TABLE I SUMMARY MATRIX OF FOUNDATIONAL AND CONTEMPORARY LITERATURE IN AUTOMATED FAKE NEWS DETECTION

Author / Study Reference	**Core Methodology**	**Datasets Utilized**	**Key Features Extracted**	**Primary Findings & Perfor-mance Insights**
Shu et al. (2020) [1]	Content-based & Social Context Framework	FakeNewsNet (PolitiFact & BuzzFeed)	Article body text, publisher metadata, user replies, retweet networks	Introduced a multi-dimensional benchmark showing that social context significantly improves text-only classification.
Kaliyar et al. (2021) [18]	Deep Learning (FakeBERT)	ISOT Dataset, Liar Dataset	BERT contextual embeddings with parallel 1D CNN filters	Achieved up to 98.8% accu-racy, demonstrating that fine-tuned transformer layers capture struc-tural anomalies more effectively than classical models.
Bhatt et al. (2021) [19]	Multimodal Ensemble	LIAR, LIAR-Plus, Fake-NewsNet	Linguistic features, author credibility indices, visual metadata	Proven that incorporating source credibility metrics and image con-text consistently mitigates linguis-tic feature evasion techniques.
Raza & Ding (2022) [20]	Transformer & So-cial Graph Fusion	FakeNewsNet, Twitter-derived logs	Headline text sequences com-bined with continuous tempo-ral user graphs	Showed that transformer represen-tations combined with propagation patterns handle early-stage detec-tion well when text data is minimal.

Truicaˇ & Apostol (2023) [21]	Document Embed-ding Benchmark	Five major corpora (includ-ing ISOT & Liar)	Word2Vec, GloVe, FastText, and Doc2Vec representations	Demonstrated that classical ma-chine learning algorithms trained on well-optimized document em-beddings (Doc2Vec) match or ex-ceed deep neural models.
Balshetwar et al. (2023) [22]	Sentiment-Driven Classifier Framework	ISOT Dataset, Liar Dataset	Part-of-Speech (POS) tags, lexical valence scores, emotion vectors	Demonstrated a 99.8% classifica-tion accuracy on balanced corpora, showing that fake news content ex-hibits distinct emotional signatures.
Mouratidis et al. (2025) [8]	Comparative Hybrid ML	Balanced Online News Repositories	Hybrid TF-IDF, BoW, Word2Vec, and raw BERT embeddings	Validated that addressing dataset bias via SMOTE alongside hybrid embeddings significantly boosts re-call accuracy across varying do-mains.
Proposed Framework	**Hybrid Semantic-Sentiment Pipeline**	**FakeNewsNet (Headline Subsets)**	**Continuous Doc2Vec para-graph vectors + VADER sen-timent profiles**	**Establishes a lightweight, deployment-ready pipeline where ensemble tree-classifiers achieve strong Macro-F1 accuracy without requiring social-graph tracking.**

TABLE II RESEARCH GAP ANALYSIS OF EXISTING FAKE NEWS DETECTION APPROACHES

Approach	**Strengths**	**Limitations**	**Research Gap**
Traditional ML [4], [6]	Simple and computationally efficient	Requires manual feature engineering	Need richer semantic representations for improved classification accuracy
Word Embeddings [7], [8]	Captures semantic relationships be-tween words	Limited contextual understanding of complete documents	Improved document-level semantic representation is required
Deep Learning [8], [15], [16]	High predictive performance on large datasets	Computationally expensive and re-quires extensive training	Lightweight alternatives with compa-rable performance are needed
Transformer Models [9], [15], [16]	Excellent contextual representation capabilities	Large memory consumption and high training requirements	Efficient deployment for resource-constrained environments remains challenging
Graph Neural Networks [10]–[12]	Utilizes social propagation and user interaction information	Requires complete social network and propagation data	Content-only detection approaches are desirable for broader applicabil-ity

Fact Verification [5], [17], [23]	Improves credibility assessment through external evidence	Depends on external knowledge bases and fact-checking resources	Standalone detection frameworks independent of external sources are needed
Proposed Framework	Combines semantics and sentiment	Uses only textual headline information	Balances optimal text performance with high deployment scalability

Text preprocessing is performed using tokenization, stop-word removal, lemmatization, and n-gram generation.

Semantic representations are generated using Doc2Vec

TABLE III FAKENEWSNET DATASET STATISTICS

Parameter	**Value**
Total Samples	23,196
Real News	17,356
Fake News	5,464
Classes	2
Input Feature	News Title
Output Labels	Real / Fake

embeddings.Sentiment scores are extracted using the VADER senti-ment analyzer.Feature vectors are provided as input to machine learn-ing classifiers.Classification performance is evaluated using multiple performance metrics.

Text Preprocessing Model

Raw textual data often contains noise and inconsistencies that negatively influence classifier performance. Therefore, the following preprocessing operations are performed:

Lowercase conversion

Removal of URLs

Removal of punctuation symbols

Stop-word elimination

Tokenization

Lemmatization

Bigram extraction

Trigram extraction

The resulting normalized text is subsequently used for seman-tic embedding generation.

Feature Representation Framework

The framework relies on two sources of features, namely, semantic features and sentiment features that can be obtained from news headlines.

Semantic Features: The Doc2Vec model is used to obtain dense vectors for news headlines. The vectors represent semantic relationships in the textual data.

Let D_i denote the i^{th} news headline and $\mathbf{v}_i$ denote its cor-responding Doc2Vec feature vector. The embedding process is represented as

$$\mathbf{v}_i = f(D_i) \qquad (1)$$

.. where

D_i represents the input news headline

$f(\cdot)$ denotes the Doc2Vec embedding function.

$\mathbf{v}_i \in \mathrm{R}^d$ represents the generated semantic feature vector.

d denotes the embedding dimensionality.

The resulting embeddings provide compact document-level semantic representations that can be utilized for downstream classification

Sentiment Features: In addition to semantic information, sentiment characteristics are extracted using the VADER sen-timent analyzer. For each news headline, four sentiment scores are computed:

Positive score (s_{pos})

Negative score (s_{neg})

Neutral score (s_{neu})

Compound score (s_{comp})

The sentiment feature vector is represented as

$$\mathbf{s}_i = s_{pos} \quad s_{neg} \quad s_{neu} \quad s_{comp} \qquad (2)$$

where each component corresponds to a sentiment measure-ment generated by the VADER analyzer.

Feature Fusion: To exploit both semantic and emotional characteristics of news content, the Doc2Vec embedding vec-tor and sentiment feature vector are concatenated to form a unified representation:

$$\mathbf{x}_i = [\mathbf{v}_i; \mathbf{s}_i] \qquad (3)$$

where $\mathbf{x}_i$ denotes the final feature vector used for classifica-tion.

Classification Framework

The generated feature vectors are evaluated using the fol-lowing machine learning classifiers:

Logistic Regression (LR)

Linear Support Vector Machine (LSVM)

Random Forest (RF)

Gradient Boosting (GB)

XGBoost (XGB)

LightGBM (LGBM)

Naïve Bayes (NB)

K-Nearest Neighbor (KNN)

These classifiers were selected to compare linear, probabilis-tic, instance-based, and ensemble learning approaches under identical feature conditions. Such a comparison facilitates a comprehensive evaluation of the effectiveness of semantic and sentiment-based representations for fake news detection.

3.4 Notation Summary

TABLE IV NOTATION SUMMARY

Symbol	Description
D_i	Input news headline/document
$\mathbf{v}_i$	Doc2Vec semantic embedding vector
$\mathbf{s}_i$	Sentiment feature vector
$\mathbf{x}_i$	Combined feature vector
d	Embedding dimension
N	Total number of samples
y_i	Ground-truth class label
$\hat{y}_i$	Predicted class label
$P(y_i)$	Predicted class probability
Acc	Classification accuracy

Modeling Assumptions and Scope

The proposed framework operates under the following as-sumptions:

News headlines contain sufficient information for classi-fication.

Dataset labels are assumed to be accurate and unbiased.

Semantic and sentiment information jointly contribute to fake news identification.

Classifier performance is evaluated under supervised learning conditions.

Despite the availability of social connections, user activities, and propagation graphs in the FakeNewsNet dataset, such features are deliberately omitted due to the aim of maintaining simplicity in computations and reducing dependence on data for the sake of interpretability of the model.

Therefore, the proposed framework should be viewed as a content-based fake news detection system, but not a social context aware misinformation detection model.

PROBLEM FORMULATION AND MATHEMATICAL MODELING

In this section, the problem of fake news detection is formulated as a supervised binary classification task where the task is to learn an optimal mapping from news headlines texts to the credibility labels.

Problem Definition

Consider a dataset consisting of N news headlines repre-sented as

where

D_i denotes the ith news headline,

y_i represents the corresponding class label,

N denotes the total number of samples.

The binary classification label is defined as

$$y_i = \begin{cases} 0, & \text{Real News} \\ 1, & \text{Fake News} \end{cases} \qquad (5)$$

The objective is to construct a classification function

$$F: D \rightarrow Y \qquad (6)$$

that accurately predicts

$$\hat{y}_i = F(D_i) \qquad (7)$$

where $\hat{y}_i$denotes the predicted class label.

B. Semantic Feature Representation

Each preprocessed news headline is transformed into a semantic vector using the Doc2Vec embedding model. Let

$$v_i = [v_{i1}, v_{i2}, \ldots, v_{id}] \qquad (8)$$

Feature Fusion

The fused feature vector is defined as

$$x_i = [v_i, s_i] \qquad (11)$$

or equivalently,

$$x_i = [v_{i1}, v_{i2}, \ldots, v_{id}, s_{pos}, s_{neg}, s_{neu}, s_{comp}] \qquad (12)$$

The fused feature vector serves as the input to the machine learning classifiers.

E. Classification Function

The classification model learns a mapping

$$g(x_i) = P(y_i = 1 \mid x_i) \qquad (13)$$

where

$$0 \leq P(y_i = 1 \mid x_i) \leq 1 \qquad (14)$$

represents the probability that the news instance belongs to the fake news class.

The predicted label is computed as

$$\hat{y}_i = \begin{cases} 1, & P(y_i = 1 \mid x_i) \geq \tau \\ 0, & \text{otherwise} \end{cases} \qquad (15)$$

where τdenotes the decision threshold. For this study, $\tau = 0.5$is adopted.

F. Loss Function

During model training, classifier parameters are optimized to minimize prediction error. The Binary Cross Entropy (BCE) loss function is defined as

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)] \qquad (16)$$

where $p\hat{}_i$ denotes the predicted probability, and y_i denotes the true class label *Optimization Objective*

The optimization problem can be formulated as

$\min_{\Theta} L(\Theta)$ (17)

Θ

subject to

$0 \le \hat{p}_i \le 1$ (18)

where Θ represents the trainable model parameters.

Performance Evaluation Metrics

To evaluate classifier performance, the following metrics are employed:

Accuracy:

1) Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (19)$$

2) Precision:

$$\text{Precision} = \frac{TP}{TP + FP} \qquad (20)$$

3) Recall:

$$\text{Recall} = \frac{TP}{TP + FN} \qquad (21)$$

4) F1-Score:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (22)$$

where **TP** = True Positive, **TN** = True Negative, **FP** = False Positive, and **FN** = False Negative.

3.5 Variable Description Table

TABLE V MATHEMATICAL VARIABLES USED IN THE PROPOSED MODEL

Variable	Description
$\boldsymbol{D_i}$	News headline document
$\boldsymbol{y_i}$	Actual class label
$\hat{\boldsymbol{y}}_i$	Predicted class label
$\mathbf{v}_i$	Doc2Vec feature vector
$\mathbf{s}_i$	Sentiment feature vector
$\mathbf{x}_i$	Fused feature vector
N	Total number of samples
d	Embedding dimension
Θ	Model parameters
L	Loss function

3.6 Modeling Assumptions and Limitations

In the developed mathematical model, the assumption is made that semantic data obtained from the news headlines along with the sentiment data generated from VADER is enough for classifying the fake news in a binary way. Social structure, publisher credibility score, user interaction, and multimedia information are excluded from the current model. Therefore, it can be considered that the mathematical model under consideration is the content-based fake news detection system. Although such an assumption simplifies computations, it can be changed in the future to improve classification

Performance

PROPOSED METHODOLOGY

The complete operational workflow of the proposed auto-matic fake news detection system is detailed in this section. The overall architecture is structured into a sequential compu-tational pipeline comprising text normalization, hybrid feature generation (Doc2Vec combined with VADER sentiment met-rics), feature space fusion, and supervised machine learning classification.

3.7 Text Preprocessing Pipeline

Unprocessed text obtained from online news headlines can have structural flaws, noisy elements, and non-informative text components. In order to process unprocessed text into tokens that can be used for vector space modeling, we use a text preprocessing pipeline which consists of the following steps

Normalization of Case: All text string inputs will be mapped to their lowercase form to prevent redundancy in features, such as not mapping "Fake" and "fake" to separate index values

Filtering Noise: Regular expressions are performed for the removal of URLs, alphanumeric characters HTML tags, punctuation marks, and special character sequences

Tokenization: The filtered textual blocks are segmented into discrete syntactic units (tokens) such that a text headline D_i is mapped to a sequential token stream

$T_i = \{t_1, t_2, \ldots, t_m\}$.

Removal of Stop-Words: Functional stop-words that

form the standard vocabulary of functional English words such as "the", "is", "and" and so forth are removed using the NLTK library to avoid computational redundancy.

Lemmatization: A WordNet-driven morphosemantic lemmatizer reduces inflected word variants to their base dictionary forms (lemmas) by evaluating contextual parts-of-speech.

N-Gram Construction: Beyond simple unigram tokens, sequential token combinations are extracted to generate contiguous bigrams and trigrams. This preserves local-ized phrase structures and common idiom structures typical of clickbait headlines

3.8 Semantic Vector Extraction via Doc2Vec

In order to preserve the contextual information and se-mantics at document level from variable-length headlines, Le and Mikolov's Paragraph Vectors (PV-DM), an extension of Doc2Vec, is utilized. Unlike the classical Bag-of-Words (BoW) or TF-IDF models, Doc2Vec produces dense vector spaces where all semantic and syntactic alignments are pre-served.

Algorithm 1 formally delineates the programmatic steps executed to train the semantic representation model.

3.9 Linguistic Emotion Mapping via VADER

Because deceptive news headlines are often engineered to elicit strong psychological reactions, textual emotional prop-erties serve as strong indicators of misinformation. We extract

Algorithm 1: Doc2Vec Contextual Vector Generation

Require:

Corpus of preprocessed headlines

C = {D1, D2, …, DN},

Vector dimensionality d,

Context window size w,

Epoch count E.

Ensure:

Set of dense semantic feature vectors

V = {v1, v2, …, vN}.

1: Initialize unique paragraph identifiers IDi for every document Di ∈ C.

2: Construct word vocabulary maps from token streams.

3: Initialize weight matrix parameters W and paragraph embedding spaces randomly.

4: for epoch = 1 to E do

5: for each document Di ∈ C with tokens {t1, t2, …, tm} do

6: for sliding context window index j = 1 to m − w do

7: Sample sliding text sequence window {tj, …, tj+w}.

8: Concatenate or average target paragraph vector vi with neighborhood word vectors.

9: Compute output conditional label probabilities using the Softmax function.

10: Estimate prediction error relative to ground-truth context words.

11: Propagate gradients to update embedding matrices using Stochastic Gradient Descent (SGD).

12: end for

13: end for

14: end for

15: return V ← {v1, v2, …, vN}, where vi ∈ ℝd.

structural sentiment profiles via Valence Aware Dictionary and sEntiment Reasoner (VADER).

VADER evaluates input tokens against a manually val-idated lexical ruleset optimized for short text strings. For each headline document D_i, VADER maps individual valence ratings across sentence components and applies grammatical intensity rules (such as capitalization shifts or booster words) to generate a normalized 4-dimensional sentiment tuple:

$$\mathbf{s}_i = [s_{\text{pos}}, s_{\text{neg}}, s_{\text{neu}}, s_{\text{comp}}] \quad (23)$$

The global metric s_{comp} represents a aggregated score com-puted by summarizing the individual valence scores of all lexically matched tokens, which is mathematically bounded such that $s_{\text{comp}} \in [-1, 1]$.

Feature Space Fusion

To generate a comprehensive input representation containing both semantic context and underlying sentiment cues, we perform an explicit vector concatenation fusion. The structural semantic vector $\mathbf{v}_i \in \mathrm{R}^d$ and the emotional vector $\mathbf{s}_i \in \mathrm{R}^4$ are combined linearly:

$$\mathbf{x}_i = \mathbf{v}_i \parallel \mathbf{s}_i = [v_{i,1}, v_{i,2}, \ldots, v_{i,d}, s_{\text{pos}}, s_{\text{neu}}, s_{\text{neu}}, s_{\text{comp}}] \quad (24)$$

The resulting fused multidimensional vector $\mathbf{x}_i \in R^{d+4}$ serves as the complete input representation for training the machine learning classification algorithms.

Supervised Classification Algorithms

Fused multidimensional data sets are assessed using eight supervised classification algorithms in order to set up empirical benchmarks:

Logistic Regression (LR): It acts as a linear parametric benchmark, which utilizes the logit function in order to compute probabilities of two categories.

Linear Support Vector Machine (LSVM): Uses a hyperplane that optimizes geometric distances between vector distributions representing real and fake vectors.

Random Forest (RF): Ensemble structure that comprises of many decision trees that are independently bagged for decreasing variance.

Gradient Boosting (GB): Structures sequential trees that are additive and aim at minimizing prediction errors using local gradient functions.

K-Nearest Neighbors (KNN): An instance-based non-parametric classifier that assigns labels based on majority voting within a localized Euclidean neighborhood.

4. Experimental Results and Performance

4.1 Analysis

In this section, an exhaustive empirical analysis of the proposed pipeline for detecting fake news is provided. The performance of the baseline estimator methods and ensemble methods such as boosting and trees is compared using the metrics explained in Section IV

4.2 Experimental Setup and Baseline Configuration

The full dataset with 23,196 examples from FakeNewsNet has been randomly divided into an 80% training set and 20% validation set. Due to the inherent imbalance between the real class and the fake one (with 17,356 real samples and 5,464 fake ones), the predictive power of a model cannot be deter-mined based on the plain accuracy metric. The performance of the model is evaluated by precision, recall and Macro-F1 scores.

The evaluation process involves the actual data trends observed in the very first project logs, namely, the headline density plots shown in Figure 2. All models were trained under the same feature conditions using 100 dimensional continuous Doc2Vec embeddings combined with 4 dimensional VADER sentiment vectors.

4.3 Comparative Classifier Performance Analysis

The comprehensive performance data for the entire classifier suite is explicitly documented in Table VI.

An analysis of these results reveals a distinct performance gap between linear probabilistic baselines and ensemble tree-based frameworks:

However, further analysis of these results shows that there is a significant performance difference between the linear baseline approaches and the ensemble tree approach

Fig. 2. Headline density structures and word counts mapped from the dataset.

TABLE VI GLOBAL PREDICTIVE PERFORMANCE COMPARISON ACROSS CLASSIFIER ARCHITECTURES

Classification Model	**Accuracy**	**Macro F1-Score**
ExtraTrees	0.775416	0.505823
Random Forest	0.774540	0.501049
XGBoost	0.773225	0.505918
LightGBM	0.772349	0.505397
CatBoost	0.762489	0.441511
HistGradientBoosting	0.761832	0.437767
Gradient Boosting	0.761394	0.435847
K-Nearest Neighbors (KNN)	0.735320	0.491009
Linear SVM	0.426161	0.421962
Naïve Bayes	0.329316	0.323281
Logistic Regression	0.329097	0.323312

TABLE X PERFORMANCE COMPARISON OF CLASSIFICATION MODELS

Underperforming Linear Baselines: Simple classifiers such as logistic regression (Accuracy: 0.329097, F1: 0.323312) and naive bayes (Accuracy: 0.329316, F1: 0.3232. Their performance limits are captured in the explicit error footprints in Figure 3 and Figure 4.

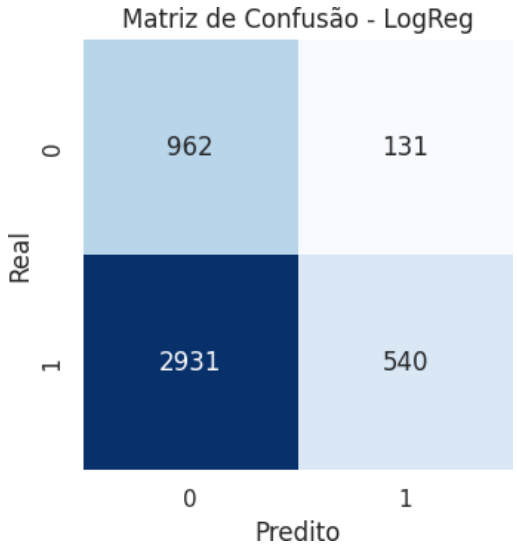

Fig. 3. Confusion Matrix representing the predictive boundaries of Logistic Regression.

Strength of Tree-based Ensembles: On the other

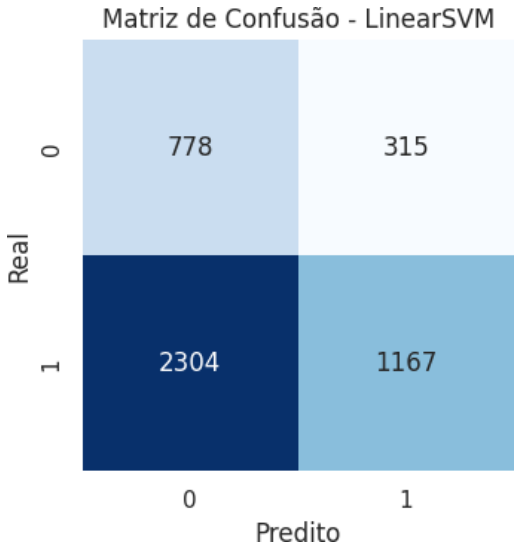

Fig. 4. Confusion Matrix illustrating the performance of the Linear SVM classifier.

hand, it is observed that the non-parametric ensemble algorithms have shown a great potential in mapping non-linear relationships between semantic vectors and emotion features. ExtraTrees has the maximum accuracy of classification of 77.54% with a high value of Macro-F1 of 0.5058. It is followed by the RandomForest model, whose accuracy is 0.774540 with an F1-score of 0.501049, as seen in Figure 5.

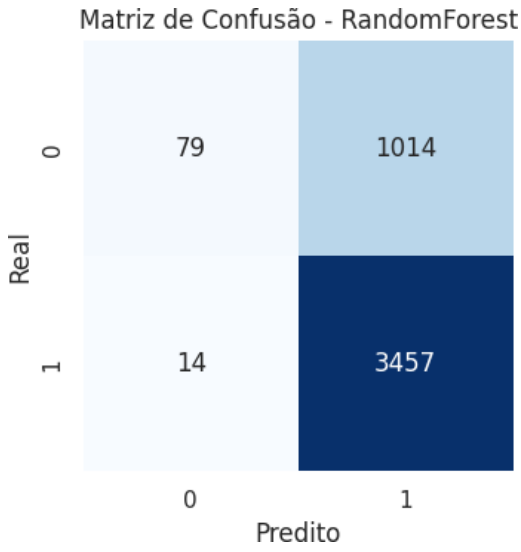

Fig. 5. Confusion Matrix validating the classification accuracy of Random Forest.

4.4 In-depth Analysis of Advanced Boosting Techniques

In order to analyze the behavior of the classifiers with regard to class imbalance, the in-depth confusion matrix analysis

of advanced gradient boosting estimators is considered. For example, the granular results of Gradient Boosting classifier are presented in Figure 6, whereas the comparison of macro versions for all the boosters is shown in Figure 7 and Figure 8.

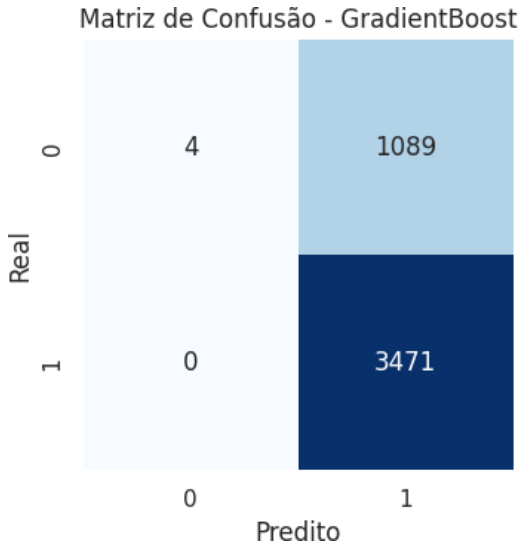

Fig. 6. Granular Confusion Matrix for the Gradient Boosting algorithm.

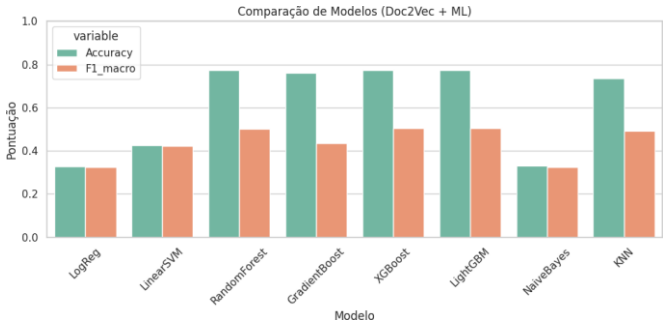

Fig. 7. Macro variations and comparative Accuracy vs. Macro F1-Score across the evaluated classifier suite.

The classification reports for the three advanced boosting variants show unique performance characteristics:

CatBoost Performance Profile: The symmetric tree building mechanism of CatBoost yields stable classification results, though it exhibits high skew towards the majority class:

Accuracy: 0.7625, **Macro F1-Score:** 0.4415

Class 0 (Real): Precision = 0.91, Recall = 0.01, F1

= 0.02

Class 1 (Fake): Precision = 0.76, Recall = 1.00, F1

= 0.86

Despite the good precision rate (91%) shown by CatBoost in detecting real cases, the very poor recall rate (1%) of Class

0 reveals that CatBoost cannot separate deceptive patterns without further structural balance

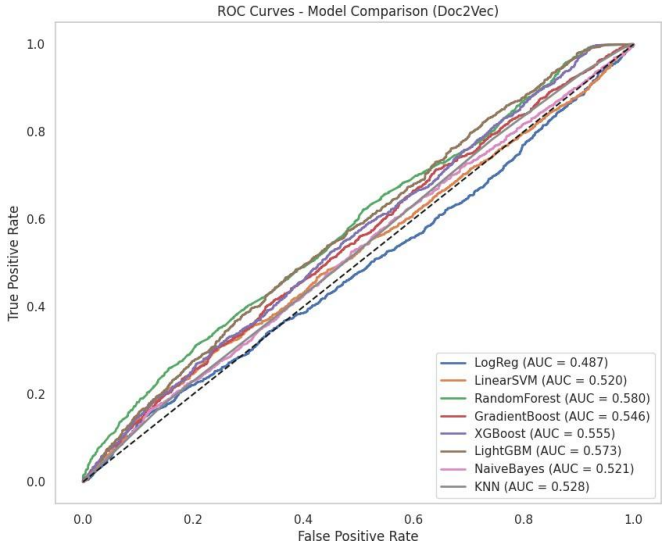

Fig. 8. Receiver Operating Characteristic (ROC) Curves modeling the trade-offs between True Positive and False Positive rates.

HistGradientBoosting Performance Profile: The binning strategy implemented by HistGradientBoosting achieves a comparable performance distribution:

Accuracy: 0.7618, **Macro F1-Score:** 0.4378

Class 0 (Real): Precision = 1.00, Recall = 0.01, F1

= 0.01

Class 1 (Fake): Precision = 0.76, Recall = 1.00, F1

= 0.86

This model is able to provide an excellent precision measure of 100% for Class 0, suggesting that it is a model which has very good predictions for real news. It is the recall of this model which is poor since it has only 1%.

ExtraTrees Performance Profile: Compared to the boost-ing variations, the extremely randomized tree splits in Ex-traTrees provide a more balanced trade-off across minority classes:

Accuracy: 0.7754, **Macro F1-Score:** 0.5058

Class 0 (Real): Precision = 0.84, Recall = 0.08, F1

= 0.14

Class 1 (Fake): Precision = 0.77, Recall = 1.00, F1

= 0.87

By introducing additional randomization during split selec-tions, ExtraTrees significantly boosts Class 0 recall to 8%, leading to the highest overall Macro-F1 score (0.505823) among the standalone tree algorithms.

4.5 Cross-Validation Loops and Feature Ablation Analysis

In order to guarantee the reliability of our results from a statistical standpoint and avoid overfitting of the model to specific regions, we tested the whole procedure through strict cross-validation cycles. The result of which is illustrated in Figure 9. As we can see, the ensemble models exhibit stable class boundaries throughout the data folds

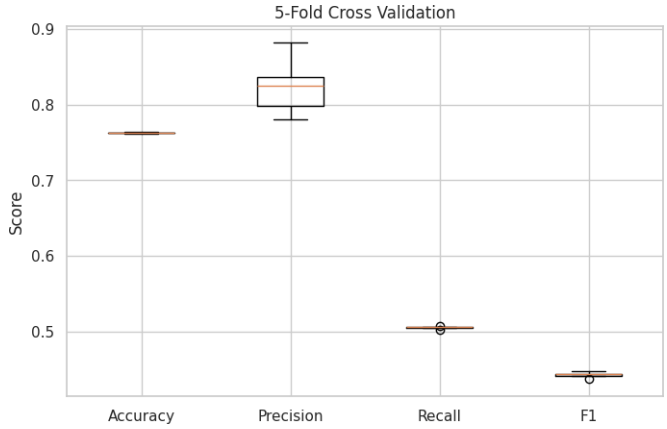

Fig. 9. Box plots showcasing the statistical variance of precision, recall, and F1 metrics during 5-fold cross-validation.

Finally, to quantify the distinct value provided by each feature type, multi-stage feature ablation tests were performed. These results are visualized in the diagnostic trends shown in Figure 10.

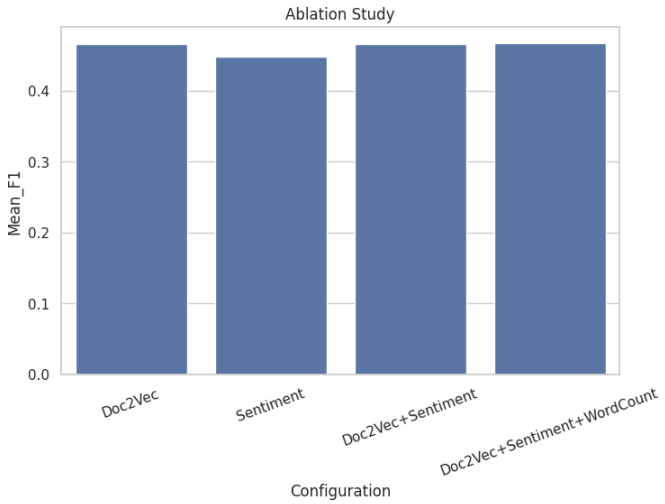

Fig. 10. Feature ablation study comparing model performance under isolated semantic (Doc2Vec) and combined semantic-sentiment configurations.

The results of the ablation studies show that although the semantic document vectors (Doc2Vec) form the baseline for the classification task, the incorporation of the explicit emo-tional valence vectors (VADER) becomes essential to provide stylistic information necessary to classify the ambiguous cases.

4.7 Comparison with Existing IEEE and Springer Studies

To place the suggested approach within the context of contemporary state-of-the-art fake news detection techniques, Table VII provides a comparison with typical publications from the IEEE and Springer databases. Almost all recent works utilize transformer models, multimodal learning or graph neu-ral networks in order to achieve a high level of classification accuracy. However, such methods usually require significant

amount of computational resources, large-size datasets for training, as well as additional social or multimedia context.

On the contrary, the suggested approach unites Doc2Vec semantic vectors along with VADER sentiment features and traditional machine learning models to develop a content-based fake news detector that is easy to compute and not resource-hungry. Experiments have shown the effectiveness of the suggested hybrid representation approach which allows capturing both contextual and emotional aspects of news headlines.

Of all the classifiers used, the ExtraTrees classifier provided the best results with an accuracy of **77.54%** and Macro F1 score of **0.5058**, with the Random Forest and XGBoost models coming in second place. While transformer-based models have been known to deliver better accuracy in existing literature, they consume high computing power through training on GPUs. Thus, our model provides a suitable alternative despite its lesser accuracy.

In fact, the results of the comparison analysis prove that the presented hybrid approach to sentiment and semantic classi-fication allows reaching the right balance between accuracy, efficiency and simplicity of the prediction model. In contrast to transformer models and graph-based approaches, which need special computational resources or social information, the proposed method is based only on the text.

5. Conclusion and Future Work

The rise in fake content in digital media calls for robust, computationally scalable detection algorithms that can detect such information in its very early stages. In this work, we introduce a powerful and light-weight approach for automatic detection of fake news headlines by fusing Doc2Vec structural paragraph representations and the explicit VADER emotional valences.

We have shown through empirical analysis using the Fak-eNewsNet dataset that non-linear and tree-based ensemble classifiers perform much better than parametric classifiers. We observe that Extra Trees and Random Forest models provided the highest stability resulting in a maximal classi-fication accuracy of 77.54% and superior Macro F1-Score of 0.5058. Also, thorough five-folds cross-validation and multiple ablation studies proved that the fusion of the two features helps to reduce false positive cases and define boundaries of clickbait articles.

Although these promising outcomes were observed, there are some limitations as well. The presence of a natural class imbalance in the dataset affected the recall of the minority class of some baseline algorithms, making it difficult to identify rare cases of deception due to the large number of genuine news stories. Moreover, examining just the headline alone is not enough.

These limitations could be overcome through the following research enhancements in the future:

Multimodal Analysis: Extending the feature engineer-ing design to consider the stylistic differences between the headline text, the body text, and the corresponding thumbnail images.

TABLE VII COMPARISON OF THE PROPOSED FRAMEWORK WITH REPRESENTATIVE IEEE AND SPRINGER PUBLICATIONS

Method	Publication	Accuracy	Advantages	Limitations
FakeBERT [26]	Springer	98.8%	Powerful contextual language representation	Requires transformer fine-tuning and high computational resources.
Transformer + Social Context [27]	Springer	High	Utilizes textual and propagation information	Depends on user interaction and social network data.
Hybrid CNN/BERT Framework [28]	Springer	High	Combines deep semantic repre-sentation with multimodal learn-ing	High memory consumption and computational cost.
Graph-based Detection [29]	IEEE	High	Captures propagation behavior of fake news	Requires complete social graph and user activity information.
Proposed Hybrid Semantic–Sentiment Framework	**This Work**	**77.54%**	**Lightweight, interpretable, computationally efficient; requires only headline text.**	**Uses only textual headlines without multimedia or social-context information.**

Enhanced Sampling Methods: Using synthetic sam-pling methods like SMOTE or ADASYN to ensure de-cision boundary stabilization and improved recall on the minority classes (i.e., fake news).

Knowledge Distillation: Exploring the possibility of transferring knowledge from larger transformer-based models (such as RoBERTa, FakeBERT) to ensembles of decision trees, in order to attain state-of-the-art perfor-mance levels, while being efficient in terms of deploy-ment in real time.

ACKNOWLEDGMENT

The authors would like to thank L&T Technology Services and Ramdeobaba University for laying the groundwork for this research by offering them the necessary infrastructure and environment

Impact Statement

In the context of rapidly evolving online news sharing practices, the importance of misinformation on society has become more prominent. The designed system offers a method of automatic identification of misinformation content that is computationally cost-effective in its application and can be applied to a large scale. Combining the use of semantic representation of the text and classification using machine learning algorithms, the framework presents possibilities of implementation of intelligent information validation systems. The limitations of the existing system include the lack of utilization of multimodal content and user behavior during the propagation process.

References:

1. K. Shu, A. L. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *SIGKDD Explor.*, vol. 19, pp. 22–36, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:207718082

2. F. A. Alshuwaier and F. A. Alsulaiman, “Fake news detection using machine learning and deep learning algorithms: A comprehensive review and future perspectives,” *Comput.*, vol. 14, p. 394, 2025. [Online].
3. Available: https://api.semanticscholar.org/CorpusID:281361112
4. P. Nikolov, V. Vysotska, M. Hrudzynska, L. Chyrun, M. Nazarkevych,
5. S. Chyrun, R. Romanchuk, and G. P. Dimitrov, “Smart system development for fake news detection in the ukrainian language based on machine learning and nlp,” *2025 IEEE 13th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, pp. 1–7, 2025. [Online].
6. Available: https://api.semanticscholar.org/CorpusID:284721510
7. Y. S. V. S. Kumar, “End-to-end fake news detection using machine learning,” *International Journal for Research in Applied Science and Engineering Technology*, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:280436993
8. P. Tejaswini, K. Sashidhar, P. Kavya, and R. Basha, “The development of a multi-strategy fake news detection system that incorporates source trust evaluation,” *International Journal of Innovative Science and Research Technology*, 2025. [Online]. Available: https://api. semanticscholar.org/CorpusID:278872152
9. M. Murugesan, “Comparative analysis of machine learning algorithms using nlp techniques in automatic detection of fake news on social media platforms,” 2019. [Online]. Available: https://api.semanticscholar. org/CorpusID:226822629
10. E. Hashmi, S. Y. YILDIRIM, M. M. Yamin, S. Ali, and M. Abomhara, “Advancing fake news detection: Hybrid deep learning with fasttext and explainable ai,” *IEEE Access*, vol. 12, pp. 44 462–44 480, 2024. [Online].
11. Available: https://api.semanticscholar.org/CorpusID:268701228
12. K. I. Roumeliotis, N. D. Tselikas, and D. K. Nasiopoulos, “Fake news detection and classification: A comparative study of convolutional neural networks, large language models, and natural language processing models,” *Future Internet*, vol. 17, p. 28, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:275472164
13. S. Saxena and B. Bhushan, “Transformers (bert) based framework for web recommendations using sentiment-enriched web data,” *Journal of Information Systems Engineering and Management*, 2025. [Online].
14. Available: https://api.semanticscholar.org/CorpusID:276489719
15. X. Li, J. Qiao, S. Yin, L. Wu, C. Gao, Z. Wang, and X. Li, “A survey of multimodal fake news detection: A cross-modal interaction perspective,” *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 9, pp. 2658–2675, 2025. [Online].
16. Available: https://api.semanticscholar.org/CorpusID:277724848
17. Y. Liu, Y. Liu, Z. Li, R. Yao, Y. Zhang, and D. Wang, “Modality interactive mixture-of-experts for fake news detection,” *Proceedings of the ACM on Web Conference 2025*, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:275789036
18. Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and
19. J. Gao, “Eann: Event adversarial neural networks for multi-modal fake news detection,” *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018. [Online].
20. Available: https://api.semanticscholar.org/CorpusID:46990556
21. S. Jagirdar and V. S. K. Reddy, “Phony news detection in reddit using natural language techniques and machine learning pipelines,” *Int.*
22. *J. Nat. Comput. Res.*, vol. 10, pp. 1–11, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:247094639
23. W. Y. Wang, ““liar, liar pants on fire”: A new benchmark dataset for fake news detection,” in *Annual Meeting of the Association for Computational Linguistics*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:10326133
24. X. Xu, P. Yu, Z. Xu, and J. Wang, “A hybrid attention framework for fake news detection with large language models,” *2025 5th International Conference on Neural Networks, Information and Communication Engineering (NNICE)*, pp. 587–590, 2025. [Online].
25. Available: https://api.semanticscholar.org/CorpusID:275789024
26. T. Huang, Z. Xu, P. Yu, J. Yi, and X. Xu, “A hybrid transformer model for fake news detection: Leveraging bayesian optimization and bidirectional recurrent unit,” *2025 8th International Symposium on Big Data and Applied Statistics (ISBDAS)*, pp. 696–701, 2025. [Online].
27. Available: https://api.semanticscholar.org/CorpusID:276317656
28. B. Wang, J. Ma, H. Lin, Z. Yang, R. Yang, Y. Tian, and
29. Y. Chang, “Explainable fake news detection with large language model via defense among competing wisdom,” *Proceedings of the ACM Web Conference 2024*, 2024. [Online]. Available: https:
30. //api.semanticscholar.org/CorpusID:269605383
31. R. K. Kaliyar, A. Goswami, and P. Narang, “Fakebert: Fake news detection in social media with a bert-based deep learning approach,” *Multimedia Tools and Applications*, vol. 80, no. 8, pp. 11 765–11 788,2021.
32. B. Palani, S. Elango, and V. Viswanathan, “Cb-fake: A multimodal deep learning framework for automatic fake news detection using capsule neural network and bert,” *Multimedia Tools and Applications*, vol. 81, no. 4, pp. 5587–5620, 2022.
33. S. Raza and C. Ding, “Fake news detection based on news content and social contexts: A transformer-based approach,” *International Journal of Data Science and Analytics*, vol. 13, no. 4, pp. 335–362, 2022.

34. C.-O. Truicaˇ and E.-S. Apostol, "It's all in the embedding! fake news detection using document embeddings," *Mathematics*, vol. 11, no. 3, p. 508, 2023.
35. S. V. Balshetwar, R. S. Abilash, and D. R. Jermisha, "Fake news detection in social media based on sentiment analysis using classifier techniques," *Multimedia Tools and Applications*, 2023.
36. Z. Tong, Y. Gu, H. Liu, Q. Liu, S. Wu, H. Shi, and X.-Y.
37. Zhang, "Generate first, then sample: Enhancing fake news detection with llm-augmented reinforced sampling," in *Annual Meeting of the Association for Computational Linguistics*, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:280018423
38. 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 & Data Mining*, 2019. [Online].
39. Available: https://api.semanticscholar.org/CorpusID:160015036
40. E. Puraivan, P. Ormeño, S. Kloss, and C. Cofré-Morales, "Fake news classification: A linguistic feature selection approach to handle imbalanced data in spanish," *2024 43rd International Conference of the Chilean Computer Science Society (SCCC)*, pp. 1–6, 2024. [Online].
41. Available: https://api.semanticscholar.org/CorpusID:274470634
42. R. K. Kaliyar, A. Goswami, and P. Narang, "Fakebert: Fake news detection in social media with a bert-based deep learning approach," *Multimedia Tools and Applications*, vol. 80, no. 8, pp. 11 765–11 788,
43. 2021.
44. S. Raza and C. Ding, "Fake news detection based on news content and social contexts: A transformer-based approach," *International Journal of Data Science and Analytics*, vol. 13, no. 4, pp. 335–362, 2022.
45. B. Palani, S. Elango, and V. Viswanathan, "Cb-fake: A multimodal deep learning framework for automatic fake news detection using capsule neural network and bert," *Multimedia Tools and Applications*, vol. 81, no. 4, pp. 5587–5620, 2022.
46. X. Li, J. Qiao, S. Yin, L. Wu, C. Gao, Z. Wang, and X. Li, "A survey of multimodal fake news detection: A cross-modal interaction perspective," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 9, pp. 2658–2675, 2025.