



A Leakage-Controlled Hybrid Word-Character TF-IDF Framework for Explainable Fake News Detection

L. Malarvizhi, MCA., M.Phil.

HOD & Assistant Professor, Department of BCA

Mohamed Sathak Hamid College of Arts & Science for Women

Ramanathapuram, Tamil Nadu, India

Email: malarhasini2013@gmail.com

Abstract

Fake news detection remains challenging because high benchmark accuracy can reflect dataset-specific stylistic or source cues rather than robust credibility assessment. This study presents a leakage-controlled hybrid text-classification framework that combines word-level and character-level TF-IDF representations with a Linear Support Vector Machine (SVM). The system was evaluated on the ISOT Fake News Dataset after cleaning, removal of exact duplicate texts, exclusion of source/subject/date metadata from predictive features, and removal of explicit Reuters tokens intended to reduce a known source shortcut. After preprocessing, 39,092 articles were retained and divided into 26,582 training, 4,691 validation, and 7,819 final held-out test instances. Four models were compared on the validation set: Logistic Regression, Linear SVM, Multinomial Naive Bayes, and the proposed Hybrid Word-Character SVM. The hybrid model achieved 99.872% validation accuracy and 99.860% F1-score and was selected before final testing. On the untouched held-out test set, the selected model achieved 99.81% accuracy, 99.83% precision, 99.75% recall, 99.79% F1-score, and 99.81% macro-F1. The implementation also maintains a model version identifier and supports integration with a desktop credibility-analysis application. These results demonstrate strong in-domain discrimination on ISOT; however, because ISOT's real and fake classes originate from systematically different source collections, the paper does not interpret the result as universal factual-verification accuracy. Cross-dataset validation and evidence-grounded fact verification are identified as essential future work.

Keywords: *fake news detection; TF-IDF; character n-grams; support vector machine; explainable AI; credibility analysis; ISOT; machine learning*

1. Introduction

The rapid circulation of misleading information through social networks, messaging platforms, blogs, and online news portals has made automatic fake news detection an important research problem. Manual fact-checking remains indispensable for high-stakes claims, but the volume and velocity of online content motivate computational systems that can prioritize suspicious material and provide decision support.

Fake news detection can be framed as supervised text classification, where a model learns statistical differences between labeled real and fake articles. Earlier approaches relied on lexical features and classical classifiers, while later research introduced neural networks, transformer encoders, social-context features, and evidence-based verification. Despite progress, benchmark performance must be interpreted carefully: a classifier may learn publisher, formatting, topic, or collection artifacts that correlate with labels without learning whether a claim is factually correct.

This work therefore focuses on two goals: obtaining a strong reproducible in-domain classifier and making the experimental protocol more defensible by reducing obvious leakage. The proposed method fuses word n-gram TF-IDF features, which capture lexical and short phrasal patterns, with character n-gram TF-IDF features, which capture subword and stylistic regularities. A Linear SVM then learns a high-dimensional separating hyperplane. The proposed model is compared with three classical baselines under the same split and preprocessing pipeline.

The main contributions are: (1) a hybrid word-character TF-IDF representation for fake-news article classification; (2) a leakage-control preprocessing procedure that excludes metadata and explicit Reuters markers and removes exact

duplicate texts before splitting; (3) a validation-based model-selection protocol followed by one evaluation on a held-out final test set; and (4) an application-oriented implementation that records model version and can be integrated into a Python/Tkinter credibility-analysis system.

2. Related Work

Wang introduced the LIAR benchmark, containing approximately 12.8K manually labeled short statements collected from PolitiFact, and demonstrated the value of combining textual and metadata information for automatic fake-news detection [1]. The benchmark also highlighted that fake-news research spans more than article-level binary classification and may involve fine-grained truthfulness labels.

The ISOT Fake News Dataset provides full news articles in separate truthful and fake collections. According to the dataset documentation, truthful articles were collected from Reuters, whereas fake articles were collected from sources identified as unreliable; most articles concern political and world news, largely from 2016-2017 [2]. This makes ISOT useful for controlled experimentation, but the class-dependent provenance creates a potential shortcut that must be acknowledged.

Ensemble and feature-fusion methods have also been used in misinformation research. Thorne et al. showed that stacked ensembles can improve over individual classifiers in stance detection [3]. Evidence-oriented work has demonstrated that incorporating justification or evidence can improve fact-checking models beyond claim-only representations [4]. More recent ISOT experiments report very high in-domain performance; for example, a 2025 Journal of Big Data study reported 99.50% accuracy for DistilBERT under its own experimental protocol [5]. Because train/test construction and preprocessing differ across studies, such published values are contextual references rather than directly comparable leader-board scores.

The present study deliberately retains a computationally efficient classical architecture. This choice provides fast training, interpretable feature engineering, modest hardware requirements, and straightforward desktop deployment, while still allowing a rigorous comparison with strong text-classification baselines.

3. Materials and Methods

3.1 Dataset

Experiments used the ISOT Fake News Dataset. The original distribution contains separate True.csv and Fake.csv files. Each article includes a title, article text, type/subject, and publication date. In the present experiment, only the title and article text were used as predictive content. After cleaning and exact-duplicate removal, 39,092 rows remained.

Dataset split: Training = 26,582 articles; Validation = 4,691 articles; Final held-out test = 7,819 articles; Total after cleaning = 39,092 articles.

3.2 Leakage-Controlled Preprocessing

The preprocessing pipeline concatenated the title and article body, normalized whitespace, and removed URLs and email-like tokens where applicable. The subject/type, source, and date fields were excluded from the predictive feature vector. Explicit occurrences of 'Reuters' were removed to reduce direct exploitation of the truthful-class source marker. Exact duplicates were detected from normalized article text and removed before the train/validation/test split. These steps do not eliminate every possible dataset artifact, but they reduce several obvious sources of leakage.

The cleaned corpus was stratified into development and final test partitions using a fixed random state. The development data were further divided into training and validation subsets. Model selection used validation macro-F1; the final test set was not used to choose the algorithm or hyperparameters.

3.3 Feature Representation

Two sparse TF-IDF representations were investigated. Word-level TF-IDF used unigram and bigram features to capture informative terms and short phrases. Character-level TF-IDF used character n-grams (3-5 characters) to capture morphology, punctuation-adjacent patterns, spelling tendencies, and stylistic signals. The proposed representation concatenates the word and character sparse matrices through feature union.

For a term t in document d , TF-IDF can be written conceptually as $\text{TF-IDF}(t,d) = \text{TF}(t,d) \times \text{IDF}(t)$, where IDF reduces the weight of terms that occur broadly across the corpus. Sublinear term-frequency scaling was used to reduce the influence of very frequent terms.

3.4 Classification Models

Four models were evaluated: TF-IDF with Logistic Regression, TF-IDF with Linear SVM, the proposed Hybrid Word-Character TF-IDF with Linear SVM, and TF-IDF with Multinomial Naive Bayes. Linear SVM is well suited to high-dimensional sparse text representations because it learns a maximum-margin decision boundary without requiring dense embeddings.

The proposed model uses the union of word and character TF-IDF features followed by a Linear SVM. This design is referred to as Hybrid_WordChar_SVM in the implementation.

3.5 Evaluation Metrics

Accuracy, precision, recall, F1-score, and macro-F1 were used. For the fake class, precision measures the proportion of predicted fake articles that are actually labeled fake; recall

measures the proportion of labeled fake articles detected by the model; and F1 is the harmonic mean of precision and recall. Macro-F1 gives equal weight to each class and is useful when evaluating balanced performance rather than relying only on overall accuracy.

4. System Implementation

The research implementation was developed in Python 3.12 using pandas for dataset handling, scikit-learn for feature extraction, model training and evaluation, joblib for model persistence, and Matplotlib for evaluation graphics. The trained model can be loaded by a Tkinter + SQLite desktop application. SQLite is used for application data such as user accounts, analysis history, source information, feedback, and audit records; the large research corpus is trained directly from CSV rather than inserted row-by-row into the GUI database.

Each trained research model receives a version identifier. The experiment reported here produced model version ISOT_20261003_113316. Versioning supports traceability by linking deployed predictions to a particular training run and evaluation record.

5. Experimental Results

5.1 Validation Comparison

Validation results — Logistic Regression: Accuracy 99.08%, Precision 99.34%, Recall 98.65%, F1 99.00%, Macro-F1 99.08%.

Linear SVM: Accuracy 99.47%, Precision 99.58%, Recall 99.26%, F1 99.42%, Macro-F1 99.46%.

Hybrid Word-Character SVM: Accuracy 99.87%, Precision 99.95%, Recall 99.77%, F1 99.86%, Macro-F1 99.87%.

Multinomial Naive Bayes: Accuracy 96.23%, Precision 94.94%, Recall 96.93%, F1 95.92%, Macro-F1 96.21%.

The proposed Hybrid Word-Character SVM produced the highest validation macro-F1 (99.87%) and validation accuracy (99.87%). It was therefore selected before the final held-out test evaluation. The character-feature branch increased training time but also improved validation performance relative to the word-only Linear SVM.

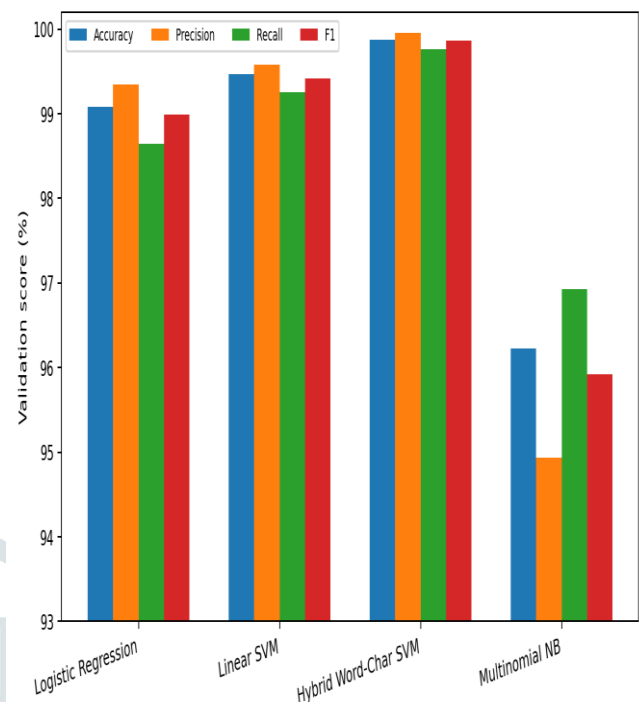


Figure 1. Validation performance of the four candidate models.

5.2 Final Held-Out Test Performance

Final held-out test — Accuracy: 99.81%; Precision (FAKE): 99.83%; Recall (FAKE): 99.75%; F1-score (FAKE): 99.79%; Macro-F1: 99.81%; Selected model: Hybrid Word-Character SVM; Model version: ISOT_20261003_113316.

The final held-out accuracy was 99.81%, with precision of 99.83%, recall of 99.75%, F1-score of 99.79%, and macro-F1 of 99.81%. These values are the measured results from the user's completed run and were not manually adjusted. The close agreement between accuracy and macro-F1 indicates strong class-balanced performance on the held-out ISOT partition.

5.3 Contextual Comparison with Published Work

Published-work context — This work: 99.81% on a leakage-reduced held-out ISOT split. A 2025 DistilBERT study reported 99.50% on ISOT under a different protocol. LIAR and the Fake News Challenge use different datasets/tasks and are not directly comparable.

The measured 99.81% is numerically higher than the 99.50% DistilBERT figure reported in the cited 2025 ISOT study, but this paper does not claim a statistically established universal superiority because the experimental protocols are not identical. A valid head-to-head superiority claim would require reimplementing competing methods on exactly the same cleaned corpus, split, leakage controls, and evaluation code.

6. Discussion

The results show that a carefully engineered sparse model can achieve extremely strong in-domain discrimination on ISOT. The improvement of the hybrid representation over word-only Linear SVM suggests that character n-grams

contribute complementary information. Character features can capture stylistic and subword patterns that are not represented explicitly by word unigrams and bigrams.

At the same time, the high result must be interpreted in light of the dataset's construction. ISOT's truthful articles were collected from Reuters while fake articles came from different unreliable sources. Consequently, even after removing explicit Reuters tokens and metadata, residual source or writing-style differences may remain. The model should therefore be described as an ISOT article classifier rather than as an independent fact-verification engine.

For real-world use, the classifier should operate as one component of a broader credibility-analysis pipeline. A practical system can combine text classification with source reputation, linguistic/clickbait indicators, uncertainty thresholds, evidence retrieval, and a 'needs verification' state. Such a design avoids presenting a statistical classifier's output as definitive proof of truth or falsity.

7. Threats to Validity and Limitations

First, the evaluation is in-domain: training, validation, and testing originate from the same dataset collection process. Second, source-specific stylistic artifacts may remain after explicit leakage reduction. Third, the model performs binary article classification and does not retrieve external evidence or verify individual factual claims. Fourth, the experiment reports a single fixed split; repeated stratified cross-validation would provide a better estimate of variance. Fifth, the study has not yet performed cross-dataset testing on LIAR, FakeNewsNet, or a newer manually curated corpus.

These limitations are important for publication. They do not invalidate the measured ISOT result, but they constrain the conclusion that can be drawn from it.

8. Conclusion and Future Work

This paper presented a leakage-controlled Hybrid Word-Character TF-IDF + Linear SVM framework for fake-news article classification. After cleaning and duplicate removal, 39,092 ISOT articles were evaluated using separate training, validation, and final held-out test partitions. The proposed hybrid model outperformed Logistic Regression, word-only Linear SVM, and Multinomial Naive Bayes on the validation set and achieved 99.81% accuracy, 99.83% precision, 99.75% recall, 99.79% F1, and 99.81% macro-F1 on the held-out test set.

Future work should evaluate the model across independent datasets and time periods, use repeated cross-validation, test transformer baselines under the same leakage-controlled protocol, and incorporate evidence retrieval and claim verification. Calibration should be added so that confidence values correspond to empirical correctness probabilities. The desktop system can also expose explanation features, source evidence, model-version

metadata, and an uncertainty-aware 'needs verification' outcome.

9. Reproducibility Statement

The experiment used a fixed random state of 42, a 20% final held-out test partition, and a validation split drawn from the development partition. The final test set was not used for model selection. The research code saves the selected model, model metadata, validation comparison table, final classification report, final results CSV, and confusion-matrix image. For publication, the authors should archive the exact code version, package versions, dataset acquisition date, and generated result files in a repository or institutional archive.

Research artifact identifier for the reported run: model version ISOT_20261003_113316. For long-term reproducibility, the exact repository or DOI should be added if the code is publicly archived.

Declarations

Data availability: The ISOT Fake News Dataset is publicly distributed by the Information Security and Object Technology (ISOT) Research Group, University of Victoria. The author complied with the dataset's applicable terms and cites the official dataset documentation.

Conflict of interest: The author declares no conflict of interest.

Funding: No external funding was reported for this study.

Ethics statement: The study uses an existing news-text dataset and does not report experiments involving human participants or animals. Institutional requirements should be checked before submission.

References

- [1] W. Y. Wang, "Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection," Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 422-426, 2017. doi:10.18653/v1/P17-2067.
- [2] Information Security and Object Technology (ISOT) Research Group, University of Victoria, "ISOT Fake News Dataset," dataset documentation, 2017/2023 distribution documentation.
- [3] J. Thorne, M. Chen, G. Myrianthous, J. Pu, X. Wang, and A. Vlachos, "Fake news stance detection using stacked ensemble of classifiers," Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism, pp. 80-83, 2017. doi:10.18653/v1/W17-4214.
- [4] T. Alhindi, S. Petridis, and S. Muresan, "Where is Your Evidence: Improving Fact-checking by Justification Modeling," Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), 2018.
- [5] "Tri-Algo guardian ensemble approach for fake news detection in social media," Journal of Big Data, 2025, reporting DistilBERT accuracy of 99.50% on ISOT under the study's protocol. doi/article identifier available from the journal page.
- [6] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.
- [7] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, pp. 273-297, 1995.