International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 6, Nov-Dec-2026, ISSN (Online): 2395-566X

Fake News Detection Using TF-IDF Features and Classical Machine Learning: A Comparative Study

Anmol Gupta[1], Suryansh Mittal[2] Vineet Mohan[3]
Cse-7a, Cse-7b Department Of Computer Science And Engineering
Hmr Institute Of Technology & Management, Hamidpur, Delhi-110036

Abstract- The rapid circulation of misleading news on digital platforms has made automated veracity assessment an urgent practical requirement. This paper presents a content-based fake news detector that classifies a news article as real or fabricated using only its text. Articles are normalised through a deterministic cleaning pipeline, represented as TF-IDF vectors over unigram and bigram features, and classified by three classical models: Logistic Regression, Multinomial Naive Bayes and a calibrated Linear Support Vector Machine. Experiments on a public corpus of 6,335 balanced news articles show that the Linear SVM achieves the highest test accuracy of 93.69%, followed by Logistic Regression at 91.79% and Multinomial Naive Bayes at 86.35%, with five-fold cross-validation confirming the same ordering at 93.05%, 91.67% and 87.26% respectively. An inspection of model coefficients shows that although attributional reporting vocabulary marks the real class, the fabricated class is separated partly by corpus-specific entity vocabulary and by website boilerplate. That finding is reported as a limitation rather than a feature: it warns that in-domain accuracy overstates real-world transfer. The study concludes that a TF-IDF pipeline of this kind remains a strong, interpretable and inexpensive baseline against which heavier transformer models should be measured.

Keywords: fake news detection; natural language processing; TF-IDF; Logistic Regression; Multinomial Naive Bayes; support vector machine; text classification; misinformation.

I. INTRODUCTION

News consumption has shifted decisively to online platforms, where publication is cheap and verification is optional. The consequence is a steady supply of fabricated stories that imitate the form of journalism, and a well-documented difficulty for readers in telling the two apart. Because manual fact-checking cannot keep pace with publication volume, computational detection has become an active research problem in natural language processing and data mining (Ahmed et al., 2017; Shu et al., 2017; Zhou and Zafarani, 2020).

Approaches in the literature fall into two families. Content-based methods examine what the article says and how it says it, using lexical, syntactic and stylistic features (Pérez-Rosas et al., 2018; Potthast et al., 2018). Context-based methods exploit auxiliary signals such as the publisher, the propagation network or user engagement on social media (Shu et al., 2017). Content-based detection is attractive in practice because it needs nothing beyond the article itself, and therefore works on text that has already been stripped of its social context.Within content-based detection, classical bag-of-words pipelines remain widely used: they are fast, need no pretrained parameters, and their decisions can be inspected. Ahmed et al. (2017) established this line of work on a large news corpus, reporting 92% accuracy with TF-IDF features and a linear SVM. Pérez-Rosas et al. (2018) later showed that stylistic cues such as punctuation and readability are strongly discriminative, while Potthast et al. (2018) found that writing style separates hyperpartisan from mainstream reporting more reliably than topic. Transformer architectures such as BERT (Devlin et al., 2019) now exceed such figures on many benchmarks, but at considerable cost in compute and interpretability, which makes the classical baseline a useful reference point rather than an obsolete one.

This work makes three contributions. First, we specify a reproducible TF-IDF pipeline for binary fake news classification. Second, we compare three classifiers under an identical feature representation and identical splits, reporting both held-out test performance and five-fold cross-validated accuracy. Third, we analyse which lexical items drive the decision, and we report the finding that part of the learned signal is corpus-specific boilerplate rather than style, which we treat as a warning about reported accuracy.

II. METHODOLOGY

1. *Dataset*

We use a publicly distributed corpus of 6,335 English news articles labelled FAKE or REAL, summarised in Table 1. The classes are almost exactly balanced, which removes the need for resampling and makes accuracy a meaningful summary statistic. The corpus is distributed through Kaggle and we obtained a verbatim copy from the public repository

© 2026 IJSRET

github.com/lutzhamel/fake-news; no cleaning beyond the pipeline described below was applied. The pipeline accepts larger corpora such as ISOT (Ahmed et al., 2017) without code changes.

Table 1. Corpus summary and experimental setup

Property	Value	Share
Total articles	6,335	100%
Labelled FAKE	3,164	49.9%
Labelled REAL	3,171	50.1%
Training / test split	5,068 / 1,267	80 / 20
TF-IDF features retained	5,000	—
Average length (fake / real)	348 / 454 words	—

After preprocessing, fake articles average 348 words and real articles 454 words, so length alone is not a reliable discriminator (Fig. 1).

2. *Text Preprocessing*

Each article is lowercased; URLs and HTML tags are removed; all punctuation and digit characters are stripped; whitespace is collapsed; and English stopwords are filtered using the scikit-learn (Pedregosa et al., 2011) stopword list. The pipeline is deliberately deterministic and parameter-free so that the comparison isolates the effect of the classifier rather than of tuning.

3. *Feature Representation*

Cleaned text is vectorised with TF-IDF, retaining the 5,000 highest-scoring unigram and bigram features and discarding terms that occur in fewer than two documents or in more than 95% of documents. Bigrams allow short discriminative phrases to be captured, while the document-frequency bounds suppress both hapax noise and uninformative high-frequency terms.

4. *Classifiers*

Three models are trained on the identical feature matrix. Logistic Regression fits a linear decision boundary with a log-loss objective and yields calibrated probabilities. Multinomial Naive Bayes models class-conditional feature counts under a strong independence assumption and serves as a fast probabilistic reference. The Linear SVM maximises the margin between classes and suits high-dimensional sparse text; because a plain linear SVM does not emit probabilities, we wrap it in a calibrated classifier using internal three-fold calibration so that confidence scores remain available at inference time.

5. *Evaluation Protocol*]

Data are split 80/20 into 5,068 training and 1,267 test articles with stratification, and the split is shared by all three models. We report accuracy, macro-averaged precision, recall and F1 on the held-out test set, plus the mean and standard deviation of accuracy over stratified five-fold cross-validation on the full corpus. Macro averaging is used because errors on either class matter equally.

III. RESULTS AND DISCUSSION

1.*Overall Performance*

Table 2 reports the results. The Linear SVM is the strongest model with 93.69% test accuracy and a macro F1 of 93.69%, closely followed by Logistic Regression at 91.79%. Multinomial Naive Bayes trails at 86.35%. The cross-validated means reproduce this ordering with small standard deviations, which indicates that the ranking is stable rather than an artefact of one favourable split.

Table 2. Test-set performance (percent) and five-fold cross-validated accuracy

Classifier	Accuracy	Precision	Recall	F1-Score	5-fold CV Accuracy
Logistic Regression	91.79	91.91	91.79	91.79	91.67 ± 0.42
Multinomial Naive Bayes	86.35	86.44	86.35	86.34	87.26 ± 1.22
Linear SVM (calibrated)	93.69	93.69	93.69	93.69	93.05 ± 0.49

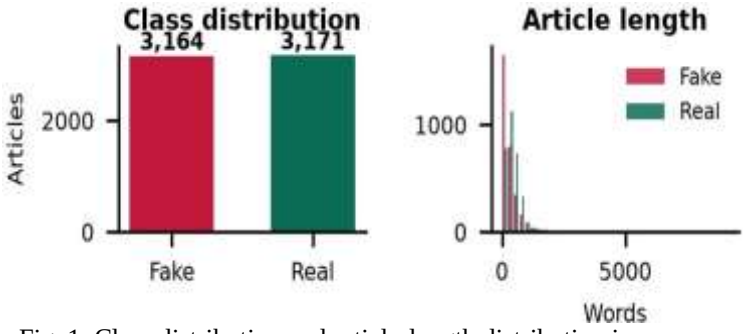

Fig. 1. Class distribution and article-length distribution in the corpus.

The margin between the SVM and Naive Bayes is about seven percentage points, which is consistent with the literature: the conditional independence assumption of Naive Bayes is poorly matched to text, where terms are strongly correlated, whereas a maximum-margin linear model tolerates correlated features well. Fig. 2 visualises the metric-

© 2026 IJSRET

by-metric comparison; the gap is largest on precision and F1, indicating that Naive Bayes produces more false positives rather than a systematic bias toward one class.

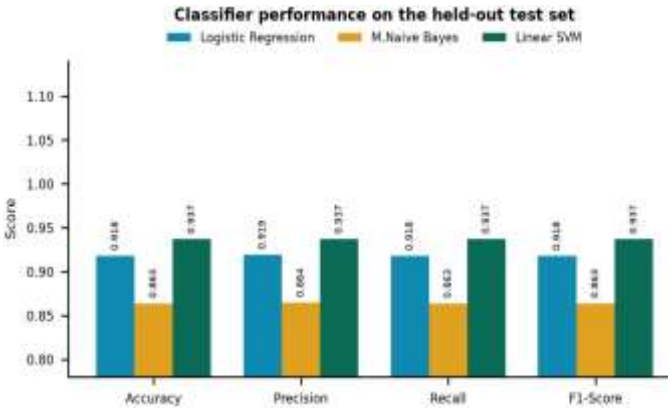

Fig. 2. Comparison of accuracy, precision, recall and macro F1 across the three classifiers.

2. *Per-Class Behaviour*

For the best model the per-class results are nearly symmetric: the FAKE class obtains precision 0.934, recall 0.940 and F1 0.937, while the REAL class obtains precision 0.940, recall 0.934 and F1 0.937. The confusion matrices in Fig. 3 confirm that errors are evenly distributed, with no tendency to collapse predictions onto the majority class.

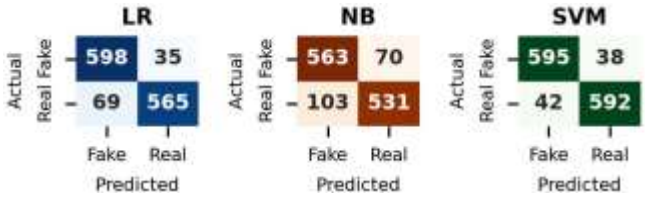

Fig. 3. Confusion matrices for LR, Naive Bayes and Linear SVM on the held-out test set.

3. Analysis Of Learned Features

Ranking the Logistic Regression coefficients yields the terms shown in Fig. 4, and the result is more sobering than a purely stylistic reading would suggest. The REAL side is dominated by attribution and campaign-reporting vocabulary, namely said, says, president, candidates, republican and debate, which is consistent with wire-service conventions. The FAKE side, however, is not dominated by sensational adjectives. It is dominated by entity vocabulary from the 2016 US election cycle (hillary, podesta, fbi, russia, election), calendar terms (october, november) and, most revealingly, website boilerplate (article, share, source). The presence of boilerplate is a warning: part of the discriminative signal separates publishing platforms rather than veracity, which inflates in-domain accuracy and is unlikely to transfer to other sources. This tempers the stylistic account of Pérez-Rosas et al. (2018) and Potthast et al. (2018), and echoes the caution in Shu et al. (2017) that content-only models can latch onto dataset artefacts.

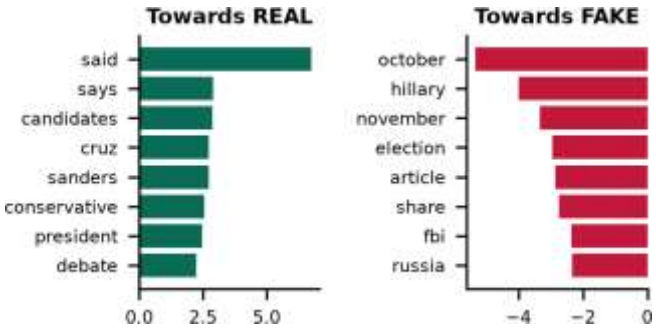

Fig. 4. Highest-weight unigram and bigram features for the two classes, from Logistic Regression.

4. Limitations Of The Study

Four limitations should be stated plainly. First, the detector assesses style, not truth: a carefully written fabrication that imitates wire-service conventions, or a genuine report that uses emotive language, will be misclassified, and the model offers no mechanism to check a claim against evidence. Second, the corpus is drawn from a particular period and set of outlets, so accuracy on a different domain or language should not be assumed; cross-domain degradation is well documented (Pérez-Rosas et al., 2018). Third, bag-of-words representations discard word order and negation, which limits sensitivity to subtle manipulation such as selective quotation. Fourth, as Section 3.3 shows, part of the learned signal is publisher-specific boilerplate, so the reported accuracy should be read as an upper bound for this corpus rather than an estimate of field performance.

These limitations do not remove the practical value of the baseline. Where the goal is triage, a model that flags likely fabrication for human review at 93.69% accuracy with millisecond inference and a fully inspectable decision is useful, and it provides the reference against which contextual and transformer-based systems should justify their additional cost.

IV. CONCLUSION

This paper compared three classical classifiers for content-based fake news detection under an identical TF-IDF representation on a balanced corpus of 6,335 news articles. A calibrated Linear SVM performed best at 93.69% test accuracy, with Logistic Regression close behind at 91.79% and Multinomial Naive Bayes at 86.35%; five-fold cross-validation confirmed the ordering. Coefficient analysis

© 2026 IJSRET

showed that the separation rests substantially on corpus-specific entity and boilerplate vocabulary, which explains the strength of the baseline in domain and predicts its weakness out of domain.

Future work will proceed along three lines: replacing TF-IDF with contextual embeddings from a transformer model (Devlin et al., 2019) and measuring the accuracy gained against the cost; adding publisher-level and metadata features to move from style toward credibility; and integrating an external fact-checking interface so that the system can support, rather than replace, human verification. We also intend to evaluate cross-domain transfer explicitly, training on one topic distribution and testing on another, since in-domain accuracy is known to overstate real-world performance.

REFERENCES

1. "Fake and Real News" corpus, 6,335 labelled articles, distributed via Kaggle; copy used from the public repository github.com/lutzhamel/fake-news (accessed November 2026).
2. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
3. H. Ahmed, I. Traore, and S. Saad, "Detecting opinion spams and fake news using text classification," Security and Privacy, vol. 1, no. 1, Wiley, 2018.
4. H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using n-gram analysis and machine learning techniques," in Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments (ISDDC 2017), LNCS vol. 10618, Springer, 2017, pp. 127–138.
5. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.
6. K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, pp. 22–36, 2017.
7. M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, "A stylometric inquiry into hyperpartisan and fake news," in Proc. 56th Annual Meeting of the Association for Computational Linguistics, 2018, pp. 231–240.
8. N. J. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news," Proceedings of the Association for Information Science and Technology, vol. 52, no. 1,pp. 1–4, 2015.
9. V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic detection of fake news," in Proc. 27th Int. Conf. on Computational Linguistics (COLING), Santa Fe, USA, 2018, pp. 3391–3401.
10. W. Y. Wang, "'Liar, liar pants on fire': A new benchmark dataset for fake news detection," in Proc. 55th Annual Meeting of the Association for Computational Linguistics, 2017, pp. 422–426.
11. X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," ACM Computing Surveys, vol. 53, no. 5, 2020.

© 2026 IJSRET