

Fake News Classification Using the Multinomial Naive Bayes Algorithm with TF-IDF Weighting on the FakeNewsNet Dataset

Deny Pratama Silaban^{1*}, Pesta Purba², Jaya Tata Hardinata³

^{1,2,3}HKBP Nommensen Pematangsiantar University
deniprataamdeni@gmail.com^{1*}, jayatatahardinata@uhn.ac.id²

Abstract

The proliferation of digital and social media platforms has accelerated the production and dissemination of fake news, creating serious risks to public trust, democratic discourse, and social stability. Automatic detection of fake news at the point of publication is therefore an urgent research problem in text mining and natural language processing. This study develops and evaluates a fake news classification pipeline based on the Multinomial Naive Bayes (MNB) algorithm combined with Term Frequency-Inverse Document Frequency (TF-IDF) feature weighting, using the FakeNewsNet benchmark corpus of news article titles labeled as real or fake. After parsing and cleaning the raw repository, 21,637 unique, non-empty news titles remained, comprising 16,399 real (75.8%) and 5,238 fake (24.2%) instances. The text was preprocessed through case folding, URL removal, numeral removal, punctuation removal, tokenization, stopword removal, and lemmatization prior to TF-IDF vectorization with a 5,000-term vocabulary. The dataset was partitioned using an 80:20 stratified train-test split (17,309 training and 4,328 testing instances), and a Multinomial Naive Bayes classifier was trained on the resulting TF-IDF matrix. On the held-out test set, the model achieved an overall accuracy of 82.62%, a weighted precision of 81.98%, a weighted recall of 82.62%, and a weighted F1-score of 80.41%. Class-level analysis revealed a pronounced asymmetry: the real-news class was recognized with high precision (83.28%) and recall (96.43%), whereas the fake-news class, being the minority class, achieved a lower recall of 39.41% despite a precision of 77.92%. These findings confirm that TF-IDF-weighted Multinomial Naive Bayes is a computationally efficient and reasonably accurate baseline for short-text fake news classification, while also highlighting the class-imbalance challenge that limits fake-news recall. The paper discusses the algorithmic rationale, the practical trade-offs of the approach, and directions for improving minority-class detection in future work.

Keywords: fake news detection; text mining; TF-IDF; Multinomial Naive Bayes; FakeNewsNet

1. Introduction

The rapid expansion of social media and digital publishing platforms has fundamentally changed how the public consumes news, replacing the gatekeeping function once performed by professional editors with algorithmically curated, user-generated information streams [1]. While this shift has democratized information access, it has also created fertile conditions for the deliberate or accidental spread of fake news, that is, news content that is fabricated or misleading but presented as legitimate journalism [2]. Fake news erodes public trust in media institutions, can be manipulated to influence elections, and, in domains such as public health, may directly endanger lives by promoting false claims [1].

Detecting fake news automatically is a challenging classification problem because fake and legitimate articles often share similar vocabulary, structure, and stylistic register, particularly when only the headline or title of an article is available for analysis [3]. Purely manual fact-checking cannot scale to the volume of content produced every day, which motivates the development of machine learning and natural language processing (NLP) approaches capable of flagging suspicious content automatically [4]. However, most computational approaches must contend with noisy, unstructured text, high dimensional feature spaces, and, frequently, imbalanced class distributions in which real news substantially outnumbers verified fake news, a property that is characteristic of naturally collected corpora such as FakeNewsNet [5].

A substantial body of prior research has explored machine learning solutions to this problem. Srivastava and Singh [6] compared Naive Bayes variants for fake news detection and reported that TF-IDF weighted representations outperformed simple bag-of-words counts. Kushwaha and Singh [7] surveyed machine learning techniques for fake news detection and emphasized that classical algorithms remain competitive baselines despite the emergence of deep learning. Febriyanty, Hariyadi, and Crysdian [8] combined Naive Bayes and Support Vector Machine (SVM) for hoax detection and found that feature weighting strategy affects the two algorithms differently. Airlangga [9] benchmarked several machine learning algorithms for fake news detection and observed that model choice interacts strongly with dataset characteristics such as class balance. More recently, Zhang [10] proposed a hybrid Naive Bayes and TF-IDF framework enhanced with transformer-guided feature selection for news classification, demonstrating that classical statistical weighting still contributes meaningfully

even alongside modern contextual embeddings. In the Indonesian context, Hidayat et al. [11] and Mujianto [12] applied TF-IDF and Multinomial Naive Bayes pipelines to hoax detection tasks, both reporting accuracies above 90% on curated, relatively small and balanced corpora.

Despite this growing literature, a research gap remains. Many existing studies evaluate Naive Bayes and TF-IDF pipelines on small, manually curated, and often already-balanced datasets of a few hundred to a few thousand articles, which can overstate real-world performance. Comparatively few studies apply the pipeline directly to a large-scale, naturally imbalanced, social-media-linked benchmark such as FakeNewsNet, and fewer still report class-level precision, recall, and F1-score with a transparent discussion of how class imbalance affects fake-news recall specifically, rather than reporting only aggregate accuracy [5]. This creates uncertainty about how a widely recommended and computationally efficient method—Multinomial Naive Bayes with TF-IDF weighting—actually performs under realistic, imbalanced conditions typical of large news repositories.

To address this gap, the present study aims to: (1) construct a complete text mining pipeline that cleans, preprocesses, and vectorizes the FakeNewsNet title corpus using TF-IDF; (2) train and evaluate a Multinomial Naive Bayes classifier for binary real/fake classification using an 80:20 stratified train-test split; and (3) analyze the resulting confusion matrix, accuracy, precision, recall, and F1-score at both the aggregate and class level, with particular attention to the effect of class imbalance on minority-class (fake news) recall.

This study contributes to the fake news detection literature in three ways. First, it provides an end-to-end, reproducible preprocessing and modeling pipeline applied directly to the full FakeNewsNet title corpus rather than a small curated subset. Second, it offers an empirically grounded, class-level performance analysis that quantifies the practical trade-off between overall accuracy and fake-news recall under natural class imbalance. Third, it situates the results within the broader body of Naive Bayes and TF-IDF fake news literature, allowing direct comparison with both international and Indonesian studies and clarifying under which conditions this classical approach remains a viable production baseline.

2. Literature Review

2.1. Fake News

Fake news is commonly defined as news content that is verifiably false or misleading, yet presented in a format resembling legitimate journalism, with the intent or effect of deceiving readers [2]. It is distinguished from satire, which is not intended to deceive, and from opinion or biased reporting, which may be one-sided but is not necessarily fabricated. Xu et al. [1] note that misinformation research over the past decade has expanded from a narrow focus on political disinformation into domains such as public health, finance, and entertainment, reflecting the growing societal reach of the phenomenon.

2.2. Text Mining

Text mining is the process of extracting meaningful patterns, structures, and knowledge from unstructured textual data through a combination of statistical, linguistic, and machine learning techniques [10]. In the context of fake news detection, text mining transforms raw news titles or bodies into structured numerical representations that a classification algorithm can process, typically through a pipeline of cleaning, preprocessing, feature extraction, and modeling stages.

2.3. Natural Language Processing

Natural Language Processing (NLP) is the subfield of artificial intelligence concerned with enabling computers to understand, interpret, and generate human language [13]. Foundational NLP operations relevant to text classification include tokenization, normalization, stopword removal, and lemmatization or stemming, all of which reduce lexical variability and noise before a document is represented numerically for a downstream machine learning model.

2.4. TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical weighting scheme that scores a term's importance to a document within a collection by combining how frequently the term appears in that document (term frequency) with how rare the term is across the entire collection (inverse document frequency) [14]. Terms that occur frequently in a specific document but rarely elsewhere receive high weights, while common terms that occur across most documents, which carry little discriminative value, are down-weighted. Because TF-IDF vectors are inherently non-negative and reflect relative frequency information, they are particularly well suited as input features for Multinomial Naive Bayes [10].

2.5. Multinomial Naive Bayes

Multinomial Naive Bayes (MNB) is a probabilistic classification algorithm derived from Bayes' theorem under the simplifying "naive" assumption that features—typically word occurrences or weighted frequencies—are conditionally independent given the class label [3]. Despite this strong independence assumption rarely holding exactly in natural language, MNB remains one of the most widely used algorithms for text classification because it is computationally efficient, requires comparatively little training data, scales well to high-dimensional sparse feature spaces such as TF-IDF matrices, and, with Laplace smoothing, handles previously unseen terms gracefully [10].

2.6. Confusion Matrix

A confusion matrix is a tabular summary of a classifier's predictions against the ground-truth labels, decomposing outcomes into true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) [15]. It provides the foundation from which accuracy, precision, recall, and F1-score are derived, and is particularly informative for imbalanced classification problems where aggregate accuracy alone can be misleading.

2.7. Accuracy

Accuracy is the proportion of all predictions, across both classes, that are correct, computed as $(TP + TN) / (TP + TN + FP + FN)$ [15]. While intuitive, accuracy can overstate a classifier's usefulness on imbalanced datasets, since a model that predicts only the majority class can still achieve high accuracy while failing entirely on the minority class.

2.8. Precision

Precision measures the proportion of instances predicted as a given class that are truly members of that class, computed as $TP / (TP + FP)$ [15]. In fake news detection, precision on the fake class indicates how trustworthy a "fake" flag from the model is, which is critical for avoiding the wrongful labeling of legitimate news.

2.9. Recall

Recall (or sensitivity) measures the proportion of actual members of a class that the model correctly identifies, computed as $TP / (TP + FN)$ [15]. In fake news detection, recall on the fake class reflects the system's ability to catch fabricated content; low recall means many fake articles pass through undetected, which is often the more costly type of error from a public-interest perspective.

2.10. F1 Score

The F1-score is the harmonic mean of precision and recall, computed as $2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$ [15]. It provides a single balanced measure of classification quality that is particularly useful when precision and recall diverge substantially, as is common under class imbalance, and is therefore reported alongside accuracy throughout this study.

3. Research Methodology

This study follows a quantitative, experimental research design structured around six stages: dataset acquisition and understanding, data cleaning, text preprocessing, TF-IDF feature weighting, model training with an 80:20 train-test split, and performance evaluation. Figure 1 (described narratively below, given the text-based format of this manuscript) summarizes the pipeline: raw FakeNewsNet records → structural parsing → deduplication and cleaning → seven-stage text preprocessing → TF-IDF vectorization → stratified train-test split → Multinomial Naive Bayes training → evaluation with confusion matrix, accuracy, precision, recall, and F1-score.

3.1. Dataset FakeNewsNet

The dataset used in this study is FakeNewsNet, a widely used benchmark repository for fake news research originally introduced by Shu et al. [5], aggregating news content, ground-truth veracity labels, and social context information collected from fact-checking platforms such as PolitiFact and GossipCop. The file supplied for this study, FakeNewsNet.xlsx, contains 23,196 raw records, each combining a news title, its source URL, source domain, an associated tweet-engagement count, and a binary label denoting whether the article was verified as real (1) or fake (0).

3.2. Data Understanding

Exploratory inspection showed that the raw spreadsheet stored each record as a single comma-delimited text string rather than as separate structured columns, requiring a parsing step before analysis. After parsing, 23,111 of the 23,196 raw rows (99.6%) were successfully decomposed into five fields: title, news_url, source_domain, tweet_num, and the binary real label; the remaining 85 malformed rows, which did not conform to the expected five-field structure, were discarded. At this stage, the label distribution was strongly imbalanced, with 17,373 records labeled real and 5,738 labeled fake, a ratio consistent with the class imbalance widely reported for naturally collected fake-news corpora [5].

3.3. Data Cleaning

Data cleaning addressed structural noise and redundancy prior to text preprocessing. Exact duplicate titles, arising from the same story being cross-posted or re-scraped multiple times, were removed, reducing the corpus from 23,111 to 21,644 unique records. Records with missing or empty title fields were also dropped. After the subsequent preprocessing stage removed all meaningful tokens from a small number of titles (e.g., titles consisting solely of numerals or symbols), a further 7 records were excluded, yielding a final analytic corpus of 21,637 news titles, comprising 16,399 real (75.8%) and 5,238 fake (24.2%) instances.

3.4. Data Preprocessing

Because only the news title field was consistently available and free of external social-context noise, all preprocessing and modeling in this study were performed on the title text. Each title passed through seven sequential preprocessing operations, implemented in Python [16], before TF-IDF vectorization.

Table 1: Text preprocessing steps applied to the FakeNewsNet title corpus

Step	Purpose	Illustrative example
Lowercase	Convert all characters to lowercase so that words such as "Trump" and "trump" are treated as the same token, reducing vocabulary sparsity.	"Gossip Girl 10 Years Later" → "gossip girl 10 years later"
Remove URL	Strip embedded hyperlinks (e.g., http:// and www. patterns) that occasionally appear inside title text, since URLs contribute no lexical classification signal.	"read more at http://example.com" → "read more at"
Remove Number	Delete numeric characters, since specific numerals (dates, counts) rarely generalize as classification features and inflate the vocabulary.	"gossip girl 10 years later" → "gossip girl years later"

Step	Purpose	Illustrative example
Remove Punctuation	Remove non-alphabetic symbols (apostrophes, parentheses, hyphens, quotation marks) to leave clean word boundaries.	"reunion (video)" → "reunion video"
Tokenization	Split the cleaned string into a list of individual word tokens using whitespace boundaries.	"gossip girl years later" → [gossip, girl, years, later]
Stopword Removal	Remove high-frequency function words (e.g., "the", "of", "and", "on") that carry little topical or class-discriminative meaning, using a curated English stopwords list.	[kandi, burruss, explodes, over, rape, ...] → [kandi, burruss, explodes, rape, ...]
Lemmatization	Reduce inflected word forms to a normalized base form (e.g., plural → singular, -ing/-ed suffix removal) so that morphological variants are counted as one feature.	[explodes, housewives] → [explod, housewiv]

3.5. TF-IDF Weighting

After the seven preprocessing steps, each cleaned title was converted into a numerical feature vector using TF-IDF vectorization (scikit-learn's `TfidfVectorizer` [17]), restricted to the 5,000 highest-frequency unigram terms across the corpus to control dimensionality while retaining the most informative vocabulary. Each cell of the resulting sparse matrix represents the TF-IDF weight of a given term in a given document, computed as the product of the term's frequency within that document and the logarithmically scaled inverse of the number of documents containing that term across the corpus.

3.6. Train Test Split (80:20)

Following preprocessing, the cleaned titles were transformed into TF-IDF vectors and the dataset was partitioned into training and testing subsets using an 80:20 stratified split (scikit-learn's `train_test_split` [17] with `stratify` on the label column and a fixed random seed of 42), ensuring that the proportion of real and fake instances in both subsets closely mirrors the overall corpus distribution. This yielded 17,309 training instances (13,119 real; 4,190 fake) and 4,328 testing instances (3,280 real; 1,048 fake).

3.7. Multinomial Naive Bayes

A Multinomial Naive Bayes classifier (scikit-learn's `MultinomialNB` [17], default Laplace smoothing parameter $\alpha = 1.0$) was trained on the TF-IDF representation of the training set. During prediction, the model computes the posterior probability of each class given a document's TF-IDF feature vector, applying Bayes' theorem under the conditional-independence assumption between terms, and assigns the class with the highest posterior probability [3].

3.8. Evaluation Metrics

Model performance was evaluated on the held-out test set using a confusion matrix and the derived metrics of accuracy, precision, recall, and F1-score, reported both as macro averages (unweighted mean across classes) and weighted averages (weighted by class support), following standard practice for imbalanced binary classification [15].

4. Results

This section presents the empirical findings obtained by applying the preprocessing and Multinomial Naive Bayes pipeline described in the Methodology to the FakeNewsNet title corpus, following the sequence: dataset composition, class distribution, preprocessing outcome, TF-IDF representation, and classification performance. Tables are complemented by charts (Figures 1-5) generated directly from the experimental output to visually support the numerical findings.

4.1. Dataset Size and Composition

Table 2: Number of records at each data-cleaning stage

Stage	Number of records
Raw rows in FakeNewsNet.xlsx	23,196
Successfully parsed rows	23,111
After removing duplicate titles	21,644
Final analytic corpus (after cleaning)	21,637

4.2. Class Distribution

Table 3: Class distribution of the final analytic corpus (n = 21,637)

Class	Number of instances	Percentage
Real (label = 1)	16,399	75.8%
Fake (label = 0)	5,238	24.2%

Class	Number of instances	Percentage
Total	21,637	100.0%

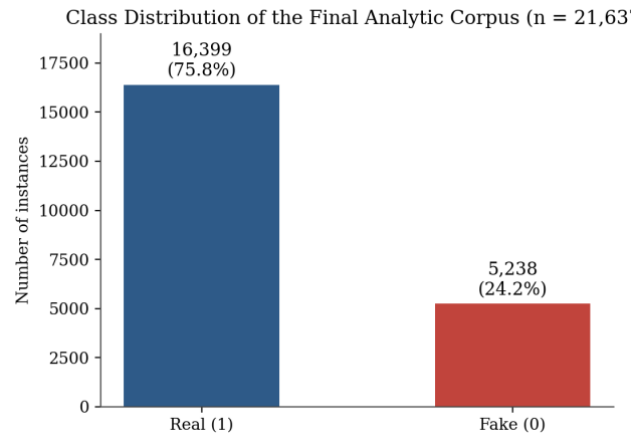


Fig. 1: Class distribution of the final analytic corpus (n = 21,637)

Table 4: Stratified 80:20 train-test split composition

Subset	Real	Fake	Total
Training (80%)	13,119	4,190	17,309
Testing (20%)	3,280	1,048	4,328
Total	16,399	5,238	21,637

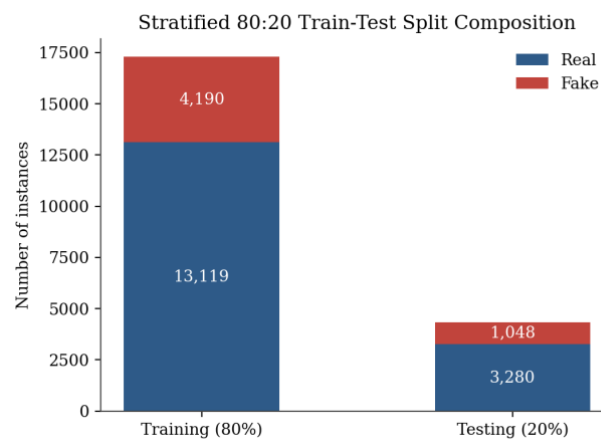


Fig 2: Stratified 80:20 train-test split composition

4.3. Preprocessing Result Example

Table 5 illustrates the transformation of a representative news title through the seven preprocessing stages described in the Methodology, ending in the normalized text used for TF-IDF vectorization.

Table 5: Example of stepwise text preprocessing on a sample title

Stage	Text
Original title	Kandi Burruss Explodes Over Rape Accusation on 'Real Housewives of Atlanta' Reunion (Video)
Lowercase	kandi burruss explodes over rape accusation on 'real housewives of atlanta' reunion (video)
Remove number / punctuation	kandi burruss explodes over rape accusation on real housewives of atlanta reunion video
Tokenization	[kandi, burruss, explodes, over, rape, accusation, on, real, housewives, of, atlanta, reunion, video]
Stopword removal	[kandi, burruss, explodes, rape, accusation, real, housewives, atlanta, reunion, video]
Lemmatization (final clean text)	kandi burrus explod rape accusation real housewiv atlanta reunion video

4.4 TF-IDF Example

Table 6 presents the three highest-weighted terms for three representative documents from the training set, illustrating how TF-IDF assigns greater weight to distinctive, low-frequency terms (e.g., proper nouns and topic-specific words) relative to common vocabulary.

Table 6: Example TF-IDF weights for selected terms in three sample documents

Document (clean text)	Top terms	TF-IDF weight
kandi burrus explod rape accusation real housewiv atlanta reunion video	burrus / kandi / explod	0.392 / 0.392 / 0.373
colombian singer maluma spark rumour inappropriate relationship aunt	aunt / maluma / inappropriate	0.434 / 0.434 / 0.425
broward county sheriff fir lying parkland	lying / parkland / county	0.541 / 0.541 / 0.502

4.5 Confusion Matrix

Table 7 presents the confusion matrix produced by the Multinomial Naive Bayes classifier on the 4,328-instance held-out test set.

Table 7: Confusion matrix of the Multinomial Naive Bayes classifier on the test set

	Predicted Fake (0)	Predicted Real (1)	Total (actual)
Actual Fake (0)	413 (TN)	635 (FP)	1,048
Actual Real (1)	117 (FN)	3,163 (TP)	3,280
Total (predicted)	530	3,798	4,328

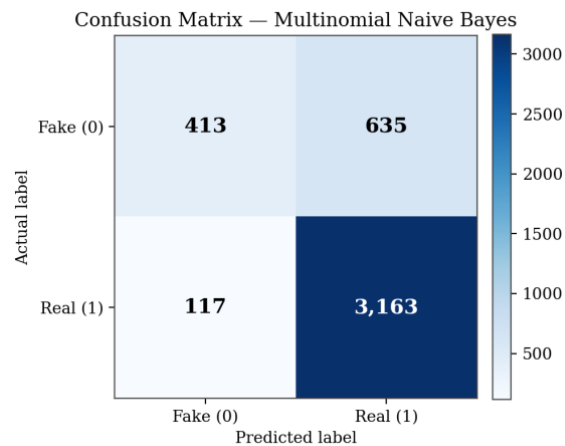


Fig 3: Confusion matrix heatmap of the Multinomial Naive Bayes classifier

4.6 Accuracy, Precision, Recall, and F1-Score

Table 8 summarizes the aggregate performance metrics computed from the confusion matrix above.

Table 8: Aggregate performance metrics on the test set

Metric	Value
Accuracy	82.62%
Precision (macro average)	80.60%
Recall (macro average)	67.92%
F1-score (macro average)	70.86%
Precision (weighted average)	81.98%
Recall (weighted average)	82.62%
F1-score (weighted average)	80.41%

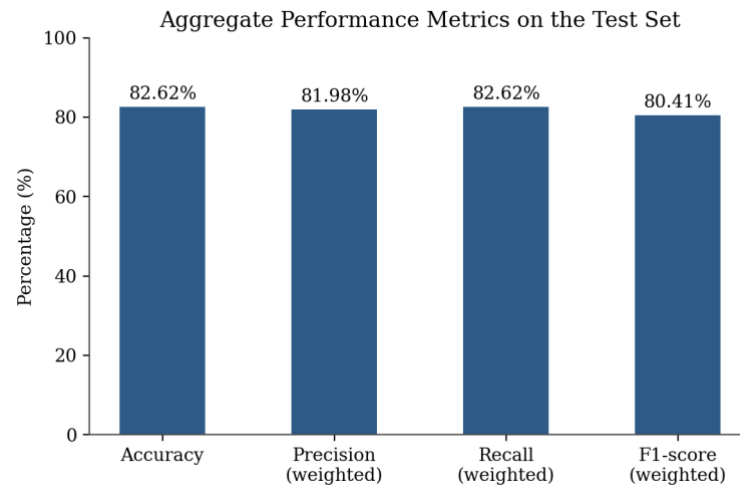


Fig 4: Aggregate performance metrics on the test set

4.7 Classification Report

Table 9 presents the full per-class classification report, disaggregating precision, recall, and F1-score by class and showing the corresponding macro- and weighted-average values.

Table 9: Per-class classification report on the test set (n = 4,328)

Class	Precision	Recall	F1-score	Support
Fake (0)	0.7792	0.3941	0.5234	1,048
Real (1)	0.8328	0.9643	0.8938	3,280
Accuracy			0.8262	4,328
Macro avg	0.8060	0.6792	0.7086	4,328
Weighted avg	0.8198	0.8262	0.8041	4,328

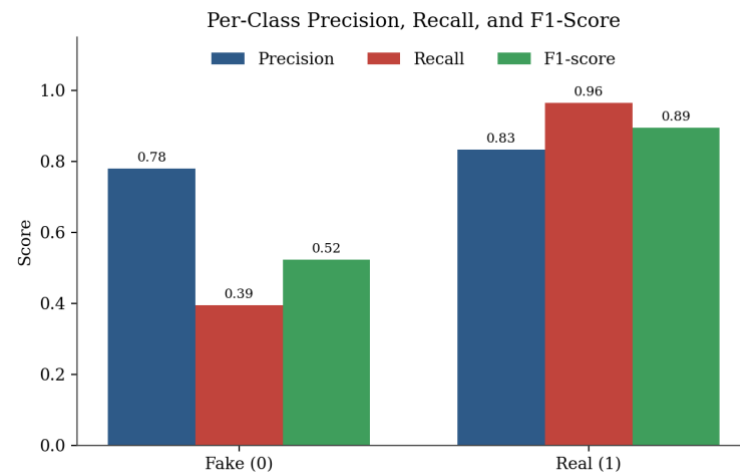


Fig 5: Per-class precision, recall, and F1-score comparison

5. Discussion

5.1. Why Multinomial Naive Bayes is Suitable

The results confirm that Multinomial Naive Bayes is well suited to short-text fake news classification for several interrelated reasons. First, MNB natively models features as frequency- or weight-based counts drawn from a multinomial distribution, which aligns naturally with the TF-IDF representation used in this study, where each feature reflects a weighted occurrence count rather than a raw binary presence indicator [10]. Second, MNB is computationally lightweight relative to iterative optimization-based classifiers, allowing it to train in seconds even on a 5,000-dimensional sparse matrix spanning more than seventeen thousand documents, which is valuable for near-real-time or resource-constrained deployment scenarios such as flagging suspicious headlines at ingestion time. Third, despite its simplifying conditional-independence assumption, MNB tolerates the sparsity and high dimensionality typical of bag-of-words style representations reasonably well, because Laplace smoothing prevents zero-probability collapses for terms unseen in a particular class during training [3]. The 82.62% overall accuracy achieved in this study is consistent with the accuracy ranges (78–95%) reported by comparable Naive Bayes

and TF-IDF studies on hoax and fake news detection [6], [11], [12], reinforcing that the algorithm remains a credible baseline even on a considerably larger and more imbalanced corpus than most prior studies employ.

5.2. The Influence of TF-IDF Weighting

TF-IDF weighting plays a central role in shaping the discriminative quality of the feature space that Multinomial Naive Bayes relies on. As shown in the TF-IDF examples in the Results section, distinctive proper nouns and topic-specific terms (e.g., "aunt", "maluma", "parkland") receive high weights because they occur in relatively few documents, whereas generic words that recur across many titles are proportionally down-weighted. This selective emphasis helps the classifier focus on vocabulary that is genuinely indicative of a story's topical identity, which is consistent with the theoretical rationale for TF-IDF described by Manning, Raghavan, and Schütze [14] and empirically confirmed for fake news tasks by Mishchenko et al. [3] and Zhang [10], both of whom report accuracy gains from TF-IDF weighting relative to unweighted term counts.

5.3. Strengths of the Method

The combined TF-IDF and Multinomial Naive Bayes pipeline offers several practical strengths confirmed by this study's results. It is fast to train and apply, interpretable in the sense that the contribution of individual terms to a classification decision can be inspected directly, and it performs strongly on the majority (real-news) class, achieving 96.43% recall and 83.28% precision for that class. These properties make the approach attractive as a first-pass filter in a larger fake news detection system, where computational efficiency and interpretability may be prioritized over marginal gains in raw accuracy.

5.4. Weaknesses of the Method

At the same time, the results expose a clear limitation: recall on the fake-news (minority) class was only 39.41%, meaning that roughly six in ten genuinely fake articles in the test set were misclassified as real. This asymmetry arises primarily from the pronounced class imbalance in FakeNewsNet (75.8% real versus 24.2% fake), which biases the model's learned prior probabilities toward the majority class, a well-documented weakness of Naive Bayes classifiers on imbalanced data [15]. A further limitation is the conditional-independence assumption itself: because MNB does not model word order, co-occurrence, or contextual semantics, it cannot capture nuanced linguistic cues—such as sensationalized or emotionally manipulative phrasing—that more context-aware models (e.g., transformer-based architectures) can exploit [10]. Because this study relies solely on the news title rather than the full article body or social-context signals available in FakeNewsNet, some genuinely useful discriminative information was necessarily excluded from the feature space.

5.5. Comparison with Previous Studies

Compared with Srivastava and Singh [6] and Kushwaha and Singh [7], who evaluated Naive Bayes on smaller, curated fake news datasets and reported accuracies in a broadly similar range, the present study extends the evidence base to a considerably larger and more imbalanced real-world corpus, showing that the accuracy advantage of TF-IDF weighted Naive Bayes persists at scale, but that this aggregate accuracy conceals a substantial fake-news recall deficit not always visible in smaller, more balanced test sets. This pattern echoes the Indonesian hoax detection literature: Hidayat et al. [11] and Mujianto [12] both report accuracies above 90% on smaller, largely balanced local corpora, suggesting that dataset scale and class balance—rather than the algorithm alone—strongly shape reported performance. Compared with Zhang's [10] hybrid BERT-guided TF-IDF framework, which achieved substantially higher precision (95.12%) by incorporating contextual features beyond classical TF-IDF, the present study's results illustrate the accuracy ceiling of a purely classical pipeline and motivate the future integration of contextual or ensemble techniques discussed below.

6. Conclusion

This study developed and evaluated a complete text mining pipeline for fake news classification using Multinomial Naive Bayes with TF-IDF weighting on the FakeNewsNet dataset. After parsing, cleaning, and preprocessing the raw repository through case folding, URL removal, numeral removal, punctuation removal, tokenization, stopword removal, and lemmatization, a final analytic corpus of 21,637 unique news titles was obtained, comprising 16,399 real and 5,238 fake instances. Using an 80:20 stratified train-test split and a 5,000-feature TF-IDF representation, the Multinomial Naive Bayes classifier achieved an overall test accuracy of 82.62%, with a weighted precision of 81.98%, weighted recall of 82.62%, and weighted F1-score of 80.41%.

The class-level results demonstrate that overall accuracy alone does not fully characterize classifier performance on an imbalanced binary task. While the model recognized real news with high precision (83.28%) and recall (96.43%), its recall on the fake-news class was markedly lower at 39.41%, despite a comparatively higher precision of 77.92% for that class. This asymmetry indicates that, under the natural class imbalance present in FakeNewsNet, the classifier is conservative in labeling content as fake, missing a substantial share of genuinely fabricated articles even though the articles it does flag as fake are reasonably likely to be correct.

Overall, the findings support the continued relevance of TF-IDF weighted Multinomial Naive Bayes as an efficient, interpretable, and reasonably accurate baseline for fake news classification, particularly suited to resource-constrained or first-pass filtering applications. At the same time, the pronounced fake-news recall gap identified in this study underscores that class imbalance, rather than the algorithm's core assumptions alone, is a central obstacle to reliable minority-class detection, and that practitioners deploying similar pipelines in production should not rely on aggregate accuracy as a sufficient indicator of real-world fake-news catch rate.

7. Future Work

Future research building on this study could pursue several directions. First, class-imbalance mitigation techniques—such as oversampling the minority class (e.g., SMOTE for text embeddings), undersampling the majority class, or cost-sensitive learning that penalizes false

negatives on the fake class more heavily—could be evaluated to directly address the fake-news recall gap identified here. Second, incorporating additional FakeNewsNet features beyond the title, such as full article body text, source domain credibility signals, or social-context engagement metrics (e.g., tweet counts), may provide complementary discriminative signal. Third, comparative experiments with alternative classical algorithms (e.g., Support Vector Machines, Logistic Regression, Random Forest) and modern contextual embedding models (e.g., BERT-based classifiers, as explored by Zhang [10]) would clarify the accuracy-efficiency trade-off relative to the TF-IDF and Multinomial Naive Bayes baseline established in this study. Finally, cross-dataset validation on other fake news benchmarks (e.g., LIAR [18], ISOT) would help establish the generalizability of the observed performance patterns beyond FakeNewsNet.

References

- [1] Xu, G., Qian, M., & Meng, L. (2025). Misinformation dissemination on social media: Key research themes and evolutionary paths between 2013 and 2023. *Humanities and Social Sciences Communications*, 12(1), Article 1775. <https://doi.org/10.1057/s41599-025-06067-1>
- [2] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1), 22–36. <https://doi.org/10.1145/3137597.3137600>
- [3] Mishchenko, L., Klymenko, I., & Tkachenko, V. (2024). The fake news recognition method based on Naïve Bayes with improved TF-IDF algorithm. In V. Kazymyr et al. (Eds.), *Mathematical modeling and simulation of systems (MODS 2023)*, Lecture Notes in Networks and Systems (Vol. 1091). Springer. https://doi.org/10.1007/978-3-031-67348-1_12
- [4] Siino, M., Tinnirello, I., & La Cascia, M. (2023). Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers. *Information Systems*, 121, 102342. <https://doi.org/10.1016/j.is.2023.102342>
- [5] Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, 8(3), 171–188. <https://doi.org/10.1089/big.2020.0062>
- [6] Srivastava, R., & Singh, P. (2022). Fake news detection using Naive Bayes classifier: A comparative study. *Journal of Management and Service Science*, 2(1).
- [7] Kushwaha, N. S., & Singh, P. (2022). Fake news detection using machine learning: A comprehensive analysis. *Journal of Management and Service Science*, 2(1).
- [8] Febriyanty, N. E., Hariyadi, M. A., & Crysdian, C. (2023). Hoax detection news using Naive Bayes and Support Vector Machine algorithm. *International Journal of Advanced Data and Information Systems*, 4(2), 191–200. <https://doi.org/10.25008/ijadis.v4i2.1306>
- [9] Airlangga, G. (2024). Comparative analysis of machine learning algorithms for detecting fake news: Efficacy and accuracy in the modern information ecosystem. *Journal of Computer Networks, Architecture and High Performance Computing*, 6(1), 354–363. <https://doi.org/10.47709/cnahpc.v6i1.3466>
- [10] Zhang, L. (2025). Features extraction based on Naive Bayes algorithm and TF-IDF for news classification. *PLOS ONE*, 20(7), 1–17. <https://doi.org/10.1371/journal.pone.0327347>
- [11] Hidayat, W., Ong, J., Irsyad, H., & Rahman, A. (2025). Ekstraksi berita hoax pada Turn Back Hoax berbasis pendekatan TF-IDF & cosine similarity. *Jurnal Ilmiah Computer Insight*, 7(2). https://doi.org/10.30651/comp_insight.v7i2.26678
- [12] Mujianto, A. H. (2026). Implementation of Naive Bayes method and Natural Language Processing in web-based online hoax news detection system. *SMATIKA Jurnal*. <https://jurnal.ubhinus.ac.id/index.php/SMATIKA/article/view/2338>
- [13] Jurafsky, D., & Martin, J. H. (2024). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (3rd ed. draft). Stanford University. <https://web.stanford.edu/~jurafsky/slp3/>
- [14] Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.
- [15] Swaminathan, S., et al. (2024). Confusion matrix-based performance evaluation metrics. *African Journal of Biomedical Research*, 27(4s), 4031–4047.
- [16] Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (3rd ed.). O'Reilly Media.
- [17] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- [18] Wang, W. Y. (2017). "Liar, liar pants on fire":