Hybrid TF-IDF and BERT Embedding Framework for Fake News Detection Using Supervised Machine Learning Models

Vathanak SENG
Faculty of Information and Computer Technologies
Paragon International University
Phnom Penh, Cambodia
vseng4@paragoniu.edu.kh

Dr.Hongly VA
Faculty of Information and Computer Technologies
Paragon International University
Phnom Penh, Cambodia
hva@paragoniu.edu.kh

***Abstract*—With the increase in the emergence of digital media and social network platforms, the production of fake news and misinformation has reached unprecedented levels. However, machine learning-based techniques for classifying fake news employ statistical methods, including TF-IDF for the representation of text which can capture text frequencies but not their semantics and context. This paper proposes a hybrid approach for feature extraction in which hybrid feature vectors are created by combining text vectors generated using the TF-IDF technique with sentence embeddings using the `all-MiniLM-L6-v2` Sentence Transformer model, and the size of the hybrid feature vector is 5,384. The five classification models used in this paper are Logistic Regression, Random Forest, XGBoost, LightGBM, and MLP. Unlike previous approaches, where a single evaluation was done using a single train/test split, the hybrid representation approach is compared to the unimodal approaches (TF-IDF and BERT) using an ablation study, and results are reported as mean $\pm$ std. deviation on five splits (seed 42-46).In case of the ISOT Fake News Dataset, the multi-modal representation reached the highest average accuracy of 98.96% $\pm$ 0.23% (MLP classifier), with a slight improvement of around 0.3 percentage points relative to the most accurate uni-modal approach (TF-IDF alone: 98.68% $\pm$ 0.16%, LightGBM classifier) and a much larger one than the BERT alone model (97.03% $\pm$ 0.25%, MLP classifier). From the ablation study results, it becomes clear that the utilization of TF-IDF features is still the leading technique for the ISOT dataset, despite the marginal improvement that the semantic information provides for the model.**

***Index Terms*—Fake News Detection, TF-IDF, BERT Embeddings, SentenceTransformer, Machine Learning, XGBoost, LightGBM, Natural Language Processing, Ablation Study**

I. Introduction

The rapid spread of misinformation via social media and online news has created new challenges for society, government, and media. The rapid and large scale dissemination of misinformation can sway public opinion toward particular political candidates or create harmful narratives in high-stake situations. Progress in natural language processing and machine learning can help turn the tides on these issues. The first generation of these solutions relied on TF-IDF vectorization combined with classifiers like logistic regression, decision trees, or random forests. While these gave reasonable results, the lack of ability to model relationships of language rendered these solutions ineffective. Later arrivals of transformer-based architectures like BERT solved this problem with contextual word representations.

This paper does not purport to be the first to unite TF-IDF with sentence embeddings for feature fusion, as combining them has already been well documented in the literature of fake news detection. We aim to rigorously measure the benefits of this fusion on the ISOT Fake News dataset by establishing strict TF-IDF-only and BERT-only baselines, and reporting results on five separate random data splits to genuinely isolate improvements from noise of a single split. Additionally, it attempts to fill the gap of previous studies that did not use ablation baselines to measure the exact contribution of the embedding and model combination of `all-MiniLM-L6-v2` when it was applied to the ISOT dataset.

- Creation of a feature extraction framework that can be reproduced, which combines sparse TF-IDF vectors and dense BERT sentence embeddings, which are then fused into a feature space of 5,384 features, with each block of features independently normalized.
- Controlled ablation for TF-IDF, BERT, and Hybrid features for each of the five supervised classifiers investigated (Logistic Regression, Random Forest, XGBoost, LightGBM, MLP).
- Five independent stratified data splits for a multi-seed approach, with Train/Validation/Test datasets to replace the typical single Train/Val/Test split, ensuring less noise in validation and reporting mean and standard deviation for all evaluation metrics.
- Estimation of the gap metric, along with training, validation, and testing accuracies, to quantify overfitting.
- Controlled hyperparameter settings to ensure no overfitting and that appropriate classification accuracy is obtained.

II. Related Work

Fake news detection has attracted significant research attention across multiple disciplines including NLP, information retrieval, and computational social science [4].

A. Traditional Machine Learning with TF-IDF

Classical classifiers combined with statistical text representation typically serve as the baseline model for fake news detection. Jouhar et al. [1] reported their findings on the combination of TF-IDF vectorization and Logistic Regression, Random Forest, Gradient Boosting, and XGBoost. They worked with a very large news dataset and report the strongest XGBoost results with the least amount of overfitting, and the results of the other ensemble methods demonstrated significantly greater amounts of overfitting.

Ahmad et al. [5] created a TF-IDF representation and LIWC2015 feature-based machine learning ensemble algorithm. They report better results using ensemble classifiers than individual classifiers. During their study, the bagging model reported the highest accuracy, while the highest precision and F1 score were reported by the XGBoost model. This motivates the use of two gradient boosting methods, XGBoost and LightGBM.

Pérez-Rosas et al. [8] studied the use of surface level linguistic features to determine news fakeness, and found that while surface level linguistic features, which are not semantic, still showed some discriminative power. Similarly, Ahmed et al. [14] were able to determine news fakeness using a combination of N-grams and some machine learning classifiers, including Linear SVM and Naive Bayes.

B. BERT-Based and Transformer Models

In this regard, the development of the BERT model [6], introduced by Devlin et al., marks the beginning of a new stage of development of NLP. Based on their research results, the usage of bi-directional representation pre-training of large unannotated corpora leads to obtaining context embedding that is highly transferable and beneficial for further tasks. As for fake news detection, the study of Kaliyar et al. [2] offers FakeBERT. It employs the pre-trained BERT embeddings in a multi-branch deep-learning model for classifying social media fake news.

Research studies on the comparison of different models have established that RoBERTa [15] is more accurate for detecting fake news than other models such as BERT or ALBERT. Even though ALBERT shows superiority over BERT in terms of computational efficiency, RoBERTa is more accurate at the expense of being resource-intensive. Hence, the decision to use a lighter `all-MiniLM-L6-v2` model [7].

In this respect, Yin and Zubiaga showed that transformer models demonstrated high sensitivity to domain shifts since they failed to detect disinformation in the health domain after training on political data [10].

C. Hybrid and Multimodal Approaches

Taking into account the consistency of statistics and semantics in text representations, numerous researchers have introduced methods of hybridization aimed at embedding the features in one vector space. In their study, Kumar et al. [3] considered deep neural networks that utilize lexical, syntactic, and structural characteristics of texts to recognize fake news. While both studies are aligned in terms of philosophy, the use of deep neural networks in Kumar et al.'s research contrasts the usage of much more concise BERT sentence encoders.

Another similar study has been performed by Aphiwongsophon and Chongstitvatana [11], where the authors combined TF-IDF representation with language features manually engineered by humans and estimated the latter using Naive Bayes, SVM, and neural networks. The feature augmentation method turned out to be successful for all classifiers used in the research, consistent with the modest hybrid gains quantified via ablation in this paper.

However, on the contrary, according to Sharma et al. [12], a multivariate technique can be used for detecting fake news, including the embeddings for text and various other features related to pictures and social graph data. While this is a move forward compared to what has been presented in this study, involving more features makes the classification process less prone to error for borderline cases where embeddings would be useful.

D. SentenceTransformers and Efficient Embeddings

SentenceTransformers is an embedding technique introduced by Reimers and Gurevych [7], which involves training transformer-based networks with a siamese and triplet method to generate semantic sentence embeddings. For our analysis, we have selected the `all-MiniLM-L6-v2` model, which produces dense 384-dimensional embeddings that maintain high semantic integrity while incurring significantly lesser inference costs compared to standard BERT encoding.

III. Dataset and Exploratory Data Analysis

A. Dataset

For this particular study, ISOT Fake News Dataset [13], [14] has been used, which is available to the public domain in the ISOT Research Lab of the University of Victoria.[1] There are two types of articles contained in this data set, that is, *fake* (label 0) and *real* (label 1). An article consists of title, body text of the article, a category, and the publication date of the article. The dataset is nearly balanced between the two classes, as shown in Fig. 1.

[1] https://onlineacademiccommunity.uvic.ca/isot/2022/11/27/fake-news-detection-datasets/

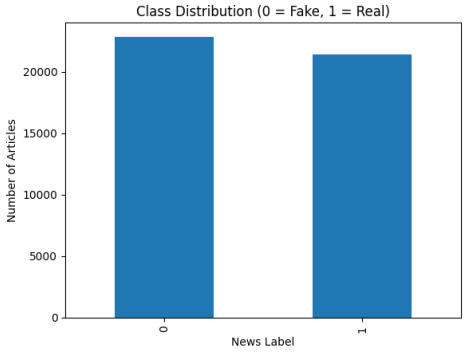


Fig. 1. Overall class distribution in the ISOT dataset: 0 = Fake, 1 = Real. The two classes are close to balanced (roughly 23,000 fake vs. 21,400 real articles).

B. Preprocessing for EDA

The following EDA was conducted prior to modelling:

- **Empty text removal.** Text rows which were empty or comprised only spaces were deleted.
- **Missing value detection.** Missing values in all the columns were checked.
- **Class distribution.** Using a bar plot, it was confirmed that the dataset is balanced in terms of fake and real news articles.
- **Subject distribution.** Article topics were visualized, and a grouped bar plot was used to observe the distribution by class.
- **Word clouds.** Separate word clouds were created to highlight common vocabularies among fake and real news article classes.
- **Data leakage prevention.** `subject`, `date`, and `title` columns were dropped before modeling to prevent data leakage. Additionally, the `source tags` of the form `CITY (Agency)` – (*e.g.*, *WASHINGTON (Reuters)* –) in the article text were eliminated using regular expression.

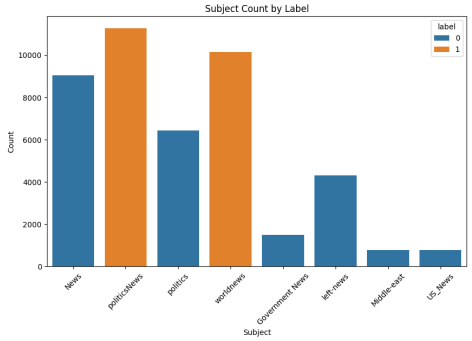


Fig. 2. Subject distribution grouped by label. Fake (0) and real (1) articles are drawn from almost entirely disjoint subject categories (e.g., *politicsNews* and *worldnews* are exclusively real; *News*, *politics*, and *left-news* are exclusively fake), motivating the removal of the `subject` column prior to modeling to avoid trivial data leakage.

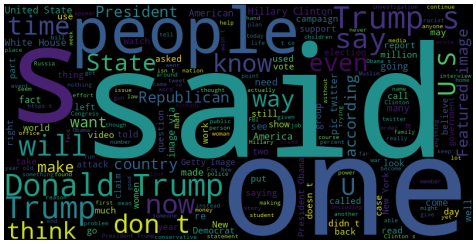


Fig. 3. Fake news (label 0) word cloud.

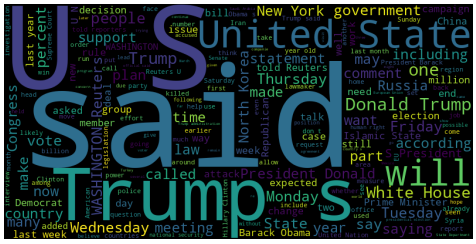


Fig. 4. Real news (label 1) word cloud. Wire-service language (*Reuters*, *WASHINGTON*) is prominent, unlike in Fig. 3.

IV. Proposed Methodology

A. Text Preprocessing

Each feature extractor works with its own distinct view of the text data (rather than a single shared view).

1) TF-IDF View: The following describes text viewed by the TF-IDF vectorizer:

Text destined for the TF-IDF vectorizer is processed as follows: Processing of text prior to TF-IDF vectorization takes place via the following steps:

- Converting to lower case
- Deleting any URLs or email addresses
- Deleting all non-alphabetic characters
- Tokenizing words using NLTK `word_tokenize` function
- Deleting stop words (English)
- Stemming with the help of Porter stemmer (words of length ≤ 2 deleted)

This aggressive normalization is appropriate for a bag-of-terms model, where reducing surface-form variation increases term-frequency signal density.

2) BERT View: Text for the SentenceTransformer encoder was lightly cleaned. In this view, only URLs and emails were removed, and the white space was adjusted. All else was left intact. Stemming and stopword removal were not applied, because `all-MiniLM-L6-v2` was trained on text that was left natural and unstemmed. Providing the Stemmed and stop word stripped text to the pipeline was in direct contradiction to the corpus the encoder was trained on, and would dramatically affect the quality of the Sentence Embeddings. With two distinct views, each feature extractor was given the appropriate level of preprocessing, and most importantly, the preprocessing that aligned with its own inductive bias.

B. TF-IDF Feature Extraction

Fitting the `TfidfVectorizer` was limited to the training split (by seed, see Section V) to prevent any leakage from the training to the test set. The settings used include:

- **max_features**: 5,000 top features
- **ngram_range**: unigrams and bigrams $(1, 2)$
- **max_df**: 0.85 (features that appear in more than 85% of documents are excluded)
- **min_df**: 5 (features that appear in less than 5 documents are excluded)
- **sublinear_tf**: True (sublinear term-frequency normalization)

The formula for the TF-IDF weight of term t in document d is

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \log\left(\frac{N}{1 + \text{DF}(t)}\right) \tag{1}$$

where N is the total number of documents and $\text{DF}(t)$ is the number of documents containing t. If `sublinear_tf=True` is used, then raw term frequency is applied to $1 + \log(\text{tf})$. In case the default ℓ_2 normalization of `TfidfVectorizer` is applied, the resulting TF-IDF vectors will be of unit length

C. Hybrid Feature Fusion

Prior to fusion, both the TF-IDF block (which is already normed to unit length) and the dense BERT block are normed independently to unit ℓ_2. None of the TF-IDF sparse block (5,000 dimensions) or the dense BERT block (384 dimensions) should dominate the concatenated representation because of the scale features. This would bias tree-based and margin-based classifiers toward the TF-IDF block, which has a higher cardinality. Once re-normalized, the TF-IDF block and the dense BERT block are concatenated horizontally with the dense BERT block converted to a sparse format (scipy.sparse.csr_matrix). Each article yields a feature vector of size 5,384.

D. Machine Learning Classifiers

Each of the five classifiers was tested on the unimodal ablation variants (VI-A) and the hybrid features. The hyper-parameters were fixed across the three feature-set variants and five random seeds.

1) Logistic Regression: $C = 0.5$ (inverse regularisation strength), `solver='lbfgs'`, `max_iter=1000`.

2) Random Forest: 100 trees, `max_depth=10`, `min_samples_leaf=5`, `max_features='sqrt'`.

3) XGBoost: 200 estimators, `max_depth=4`, `learning_rate=0.05`, `subsample=0.7`, `colsample_bytree=0.7`, L1 regularisation $\alpha = 0.1$, L2 regularisation $\lambda = 1.5$, `min_child_weight=5`.

4) LightGBM: 300 estimators, `num_leaves=63`, `max_depth=8`, `learning_rate=0.05`, `min_child_samples=20`, $\alpha = 0.05$, $\lambda = 0.3$, `subsample=0.8`, `colsample_bytree=0.8`.

5) Multi-Layer Perceptron (MLP): Three hidden layers of sizes $(256, 128, 64)$, L2 penalty $\alpha = 0.01$, early stopping enabled with `validation_fraction=0.1` and patience of 10 iterations, `max_iter=200`.

V. Experimental Setup

A. Data Splits

The dataset was split into stratified training (80%), validation (10%), and test (10%) subsets with respect to the target label. The TF-IDF vectorizer was fit on the training set only, and the vocabulary and IDF weights were transferred to the validation and test sets.

B. Multi-Seed Evaluation

To address the concern that a single data split may not reliably distinguish genuine performance differences from split-specific noise, the whole split-fit-evaluate procedure was performed independently for five random seeds ($\{42, 43, 44, 45, 46\}$), each yielding a novel stratified 80/10/10 partition. Because they do not rely on the partition, the BERT embeddings (Section IV-C) are computed once and sliced for each split. The TF-IDF vectorizer, which is only trained on the TFIDF train set, is fitted for each seed. All the metrics reported in Section VI are the mean and standard deviation over the seed-specific test set results.

C. Ablation Design

To examine the contribution of hybrid fusion, three feature-set variants were analyzed in the same way:

- **TF-IDF only**: the ℓ_2-normalized 5,000-dimensional TF-IDF vector.
- **BERT only**: the ℓ_2-normalized 384-dimensional sentence embedding, computed on the raw-text view (Section IV-C).
- **Hybrid**: the 5,384-dimensional concatenation of both normalized blocks (Section IV-D).

D. Evaluation Metrics

To evaluate classification, the following metrics were calculated based on results from the held-out test set:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

ROC-AUC was calculated based on the estimated positive-class probability (or decision score, for cases where `predict_proba` was unavailable) in relation to the true labels.

E. Overfitting Diagnosis

As a preliminary single-split (seed $= 42$) assessment that was performed prior to multi-seed testing, the gap between training and testing accuracy was calculated for each model with the hybrid feature set:

$$\text{Gap} = \text{Acc}_{\text{train}} - \text{Acc}_{\text{test}} \tag{5}$$

Models were considered to be in the "Good fit" (Gap ≤ 0.02), "Slight overfit" ($0.02 < \text{Gap} \leq 0.05$), or "Overfitting" (Gap > 0.05) categories. This single-split assessment was kept as additional evidence. The primary evidence for generalization in this revision was the low across-seed standard deviation displayed in Table I.

VI. Results and Discussion

A. Ablation Study: TF-IDF-only vs. BERT-only vs. Hybrid

Table I reports mean $\pm$ standard deviation test accuracy across five seeds for each of the three feature-set variants and all five classifiers.

TABLE I
Ablation: Mean Test Accuracy $\pm$ Std over 5 Seeds

Model	TF-IDF only	BERT only	Hybrid
Logistic Regression	0.9802±.0022	0.9334±.0032	0.9827±.0020
Random Forest	0.9594±.0075	0.8965±.0022	0.9550±.0040
XGBoost	0.9617±.0047	0.9172±.0024	0.9730±.0018
LightGBM	0.9868±.0016	0.9508±.0023	0.9891±.0010
MLP Classifier	0.9863±.0016	0.9703±.0025	**0.9896±.0023**

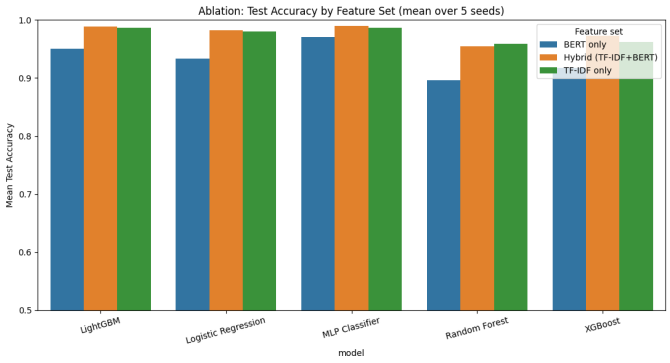


Fig. 5. Ablation comparison: mean test accuracy $\pm$ std (5 seeds) for TF-IDF-only, BERT-only, and Hybrid feature sets, across all five classifiers.

B. Full Multi-Seed Performance: Hybrid Representation

Table II reports the complete set of metrics (mean $\pm$ std over 5 seeds) for the hybrid representation, replacing the single-split results of the original evaluation.

TABLE II
Test-Set Performance, Hybrid TF-IDF+BERT
(Mean $\pm$ Std over 5 Seeds)

Model	Acc.	Prec.	Rec.	F1	AUC
Log. Reg.	.9827	.9779	.9866	.9822	.9983
Rand. Forest	.9550	.9519	.9556	.9537	.9917
XGBoost	.9730	.9671	.9776	.9723	.9966
LightGBM	.9891	.9872	.9904	.9888	.9992
MLP	**.9896**	.9891	.9894	**.9893**	.9990

Standard deviations across the 5 seeds are all ≤ 0.004 for every metric and model; see supplementary results file for per-metric std values.

TABLE III
Preliminary Overfitting Diagnosis (Single Split, Seed=42, Hybrid Features)

Model	Train	Val	Test	Gap
Logistic Regression	0.9831	0.9762	0.9801	0.0030
Random Forest	0.9688	0.9531	0.9489	0.0199
XGBoost	0.9780	0.9683	0.9699	0.0081
LightGBM	0.9999	0.9866	0.9862	0.0137
MLP Classifier	0.9995	0.9882	0.9900	0.0094

C. Discussion

The ablation results in Table I provide more insight and a deeper truth than a hybrid-only assessment. Three points are especially notable.

TF-IDF provides a solid baseline. LightGBM and MLP achieve 98.68% and 98.63% mean accuracy, respectively. This aligns with literature stating ISOT is a benchmark with a low level of difficulty. Distinction between fake and real articles is evident through their source styles as ISOT articles draw from different (Reuters vs. flagged) avenues and thus, differ in styles.

BERT fails to provide an advantage over TF-IDF. This is true for all classifiers, and is very apparent in the case of the Random Forest (89.65% vs. 95.94%) and XGBoost (91.72% vs. 96.17%) classifiers. On ISOT, this suggests that the principal clues are lexical in nature and provide a stylistic approach rather than meaning. A single 384-dimensional sentence embedding per article discards much of the information that TF-IDF's 5,000-dimensional sparse representation retains.

The hybrid representation builds on TF-IDF by adding some additional utility. For the two most notable classifiers, the hybrid representation gain over the TF-IDF-only baseline is small, but for all seeds, LightGBM displays an average accuracy of 98.91%±0.10% (an improvement over 98.68%±0.16%); MLP displays an average accuracy of 98.96%±0.23% (an improvement over 98.63%±0.16%). The hybrid representation also provided the largest gain for XGBoost (96.17% to 97.30%). Given the non-overlapping, or almost overlapping, standard deviations, this gain is not likely due to split noise; however, the improvement should not be characterized (as the original submission stated) as large, given the improvements were small. We retract that statement and

say hybrid fusion helps with small improvements most of the time. The hybrid fusion representation should be viewed as adding a complementary signal along a strong lexical baseline, and not a primary signal for driving performance on ISOT.

With Random Forest, the hybrid representation performed even *worse* than TF-IDF, as described in the original submission (95.50% vs. 95.94%). This likely means that concatenating the dense BERT block with a TF-IDF split of `max_features='sqrt'` is diluting the tree-splitting signal for this particular ensemble. LightGBM and XGBoost did not display this diluting effect, due to the leaf-wise growth and boosting corrections, respectively.

We have kept the single-split overfitting diagnosis from the original submission as a complementary, illustrative check. The low standard deviations from Tables I and II (all ≤ 0.4 percentage points) are now the most notable evidence supporting that the performance of the classification models is stable and not an artifact of a single, perhaps overly beneficial, split.

The MLP Classifier shows the best mean hybrid test accuracy with 98.96% and F1 score with 98.93%. Applying early stopping (patience 10) and L2 regularization ($\alpha = 0.01$) was effective against the overfitting of the high-dimensional 5,384-feature hybrid input, supported by the low single-split gap (0.0094) in Table III. LightGBM places second and has the highest AUC among all models (0.9992) due to its leaf-wise tree growth configuration (`num_leaves=63`), which likely captures the high-dimensional hybrid feature space effectively.

VII. Conclusion and Future Work

This paper presents a hybrid method to detect fake news using sparse TF-IDF statistics and dense BERT sentence embeddings via the lightweight `all-MiniLM-L6-v2` SentenceTransformer, with each block constructed on a preprocessing step catering to its own inductive bias, and then independently normalized prior to fusion. Different from a single-split evaluation, the hybrid representation was validated using a controlled ablation against the TF-IDF-only and the BERT-only baselines, and was assessed over five independent, stratified splits. The hybrid representation achieved the best mean test accuracy of 98.96% (MLP Classifier, $\pm$ 0.23%), and a small yet consistent gain of 0.3 percentage points on average over the best unimodal baseline (TF-IDF-only). This finding suggests that in the ISOT dataset, the dominant signal stems from the lexical features, while the semantic sentence embeddings serve a complementary role.

Future work could explore:

- Finetuning the BERT encoder in an end-to-end manner on the task of fake news classification, instead of using inference-only pre-computed embeddings, to find if task-specific finetuning improves performance beyond the modest gain observed here.
- Using SHAP to perform a feature importance analysis to determine the top informative features based on the TF-IDF values and BERT embeddings to enhance interpretability and trust in the models.
- Multilingual testing of the hybrid approach using the XLM-RoBERTa model to test its generalizability outside English datasets.
- Cross-dataset evaluation (e.g., training on ISOT and testing on FakeNewsNet [9]), to assess whether the hybrid representation's modest advantage over TF-IDF-only generalizes beyond ISOT's comparatively saturated, stylistically-separable setting.

References

[1] J. Jouhar, A. Pratap, N. Tijo, and M. Mony, "Fake news detection using Python and machine learning," *Procedia Computer Science*, vol. 233, pp. 763–771, 2024.

[2] 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. 11765–11788, 2021.

[3] S. Kumar, R. Asthana, S. Upadhyay, N. Upreti, and M. Akbar, "Fake news detection using deep learning models: A novel approach," *Transactions on Emerging Telecommunications Technologies*, vol. 31, no. 2, e3767, 2020.

[4] C. 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.

[5] I. Ahmad, M. Yousaf, S. Yousaf, and M. O. Ahmad, "Fake news detection using machine learning ensemble methods," *Complexity*, vol. 2020, article ID 8885861, 2020.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Minneapolis, MN, USA, pp. 4171–4186, 2019.

[7] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. EMNLP*, Hong Kong, China, pp. 3982–3992, 2019.

[8] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea,"Automatic detection of fake news," in *Proc. COLING*, Santa Fe, NM, USA, pp. 3391–3401, 2018.

[9] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," *Big Data*, vol. 8, no. 3, pp. 171–188, 2020.

[10] W. Yin and A. Zubiaga, "Towards generalisable fake news detection: A data resampling approach," *PeerJ Computer Science*, vol. 7, e816, 2021.

[11] S. Aphiwongsophon and P. Chongstitvatana, "Detecting fake news with machine learning method," in *Proc. ECTI-CON*, Chiang Rai, Thailand, pp. 528–531, 2018.

[12] P. Sharma, S. Sharma, S. Singh, and R. Chandra, "Combating fake news: A survey on identification and mitigation techniques," *ACM Transactions on Intelligent Systems and Technology*, vol. 12, no. 3, pp. 1–42, 2021.

[13] H. Ahmed, I. Traore, and S. Saad, "Detecting opinion spams and fake news using text classification," *Journal of Security and Privacy*, vol. 1, no. 1, Jan./Feb. 2018.

[14] 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*, I. Traore, I. Woungang, and A. Awad, Eds., Lecture Notes in Computer Science, vol. 10618, pp. 127–138, Springer, Cham, 2017.

[15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.