
A Hybrid Meta-Learning Framework for Robust Fake Review Detection Using KNN, GMM, and BERT Embeddings

Maniyarasan V and Senthilkumar G

*Department of Computer Applications, Periyar Maniammai Institute of Science and
Technology, Tamilnadu, India, gsrtpj@gmail.com and Assistant Professor -SG/CA,
Periyar Maniammai Institute of Science & Technology, Tamilnadu, India,
senthil@pmu.edu*

Abstract.

The validity of customer reviews plays an important role in ensuring consumers' trust in e-commerce websites because there are increasing numbers of artificial reviews that jeopardize online marketplaces. In this paper, a meta-learning-based hybrid model named Hybrid-XGB for fake reviews detection through three-level stacked ensemble is introduced. Sentiment polarity, length, and TF-IDF based features are generated and then processed with KNN and Gaussian Mixture Model (GMM) with linear combination, while fine-tuned BERT generates contextual and semantic features. The outputs from all base models are fed into the meta-classifier named XGBoost and combined in a meta-feature vector. Hybrid-XGB reaches 98.47% accuracy on the Amazon product reviews dataset, with both precision and recall over 98%.

Keywords. fake review detection, opinion spam, BERT, KNN, GMM, XGBoost, stacked ensemble, meta learning, TF-IDF, e-commerce.

1. INTRODUCTION

Several recent studies have noted the fast development of techniques used in fake reviews detection, stressing the move from machine learning to transformers and hybrid techniques. Also, several key problems like the lack of labeled data, adversarial reviews, and poor domain-generalization have been identified [1]. A framework based on linguistic exaggeration, BERT representations, and fuzzy logic for detecting fake reviews was proposed. Fuzzy inference allowed improving the interpretability and robustness of the method [2]. A deep learning framework using BERT for context-based representation of language allowed for a significant improvement in the detection of deceptive reviews compared to traditional feature-based methods [3]. A new hybrid GPT-BERT framework allowed combining the generative and discriminative powers of both algorithms and obtaining better results in terms of semantic feature extraction and classification [4]. The use of BERT to learn richer representations for detecting fake reviews was proposed. The technique proved to be more effective than traditional deep learning methods [5].

2. RELATED WORK

The researchers from Tao et al. (2019) have suggested the concept of a Multi-View Attention Network (MVAN) that detects fraudulent real-money trading through the use of an attention-based integration of several behavior views [6]. Gu et al. (2019) developed the concept of Interactive Matching Network for the selection of multi-turn responses in retrieval-based chatbots. They considered contextual dependencies during the communication process, improving semantic matching and response selection accuracy [7]. Ceska et al. (2020) presented an adaptively verifiability-driven strategy of the evolutionary approximation of arithmetic circuits that ensures the efficient optimization of the problem and circuit reliability to produce the optimal solutions [8]. The authors from Sanz et al. (2021) have developed a fuzzy association rule-based classifier that allows solving imbalanced classification problems. The authors have managed to improve minority class prediction by combining fuzzy logic and association rule mining [9]. Finally, the study of Jindal and Liu (2008) is the first research paper related to the issue of fake review detection and opinion spam analysis [10].

3. PROPOSED METHODOLOGY

The proposed system comprises five sequential stages: data acquisition and preprocessing, heuristic label generation, multimodal feature engineering, parallel base-model training, and XGBoost-based meta-learning fusion. Fig. 1 presents the complete architectural flowchart of the proposed system.

3.1. Dataset and Preprocessing

Experiments were performed on the publicly available Amazon product review data set (amazonsales2025INR.csv) that contains the reviews of customers with star ratings and product metadata. The data set consists of reviews extracted from multiple product domains and is rich in text and rating features that can be used to detect the fake reviews. Heuristic labeling scheme, post preprocessing of the data set, was used to generate binary classes labels (Fake and Genuine) for the supervised learning task. This data set was split into train, validation, and test data sets in the ratio of 70:15:15. This is due to the varied nature of review texts and their associated metadata and therefore constitutes a suitable dataset for testing the efficiency of the Hybrid-XGB model.

3.2. Heuristic Label Generation

Ground truth labels are seldom provided for detecting fake reviews. So, a labeling procedure based on the behavior-based approach developed by previous research [10, 25] has been employed to create a supervised training dataset. The labels "fake" (1) have been assigned to the reviews when the rating was ≤ 2 with less than 40 characters in the text, or the rating was 5 with less than 15 characters in the text. All other reviews have received the label "genuine" (0).

3.3. Feature Engineering

Linguistic Characteristics: Polarity of the sentiment is measured using TextBlob in the range from -1 to 1. Review length and character number are also taken into consideration because fake reviews tend to be shorter than real ones. TF-IDF Vectorization: The TF-IDF

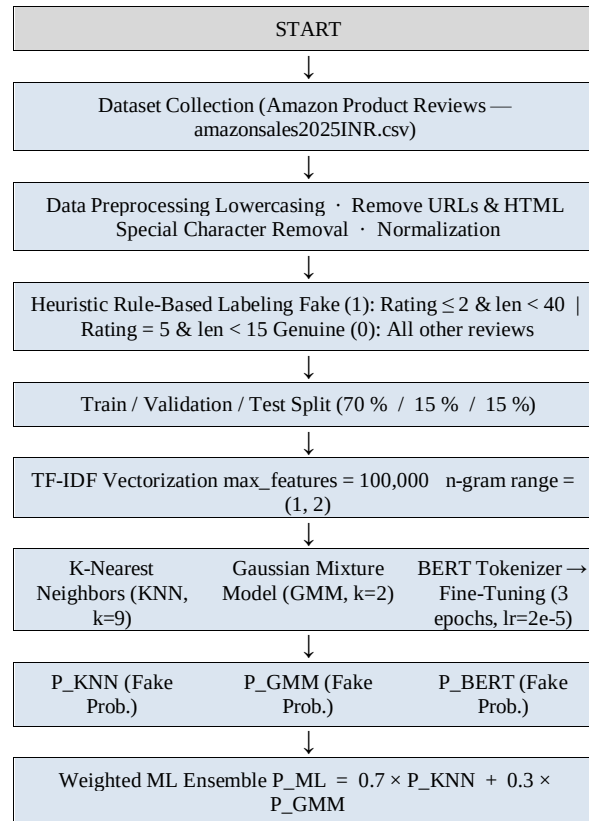
vectorizer has 100,000 maximum features and 1,2 uni-and bigrams to identify the statistical significance of words and two-word expressions.

3.4. Base Model Training

BERT has been fine-tuned through the use of the Hugging Face Trainer API, where the batch size is set to 16 and the maximum length of sequence is 128. The AdamW optimizer has been used, whereby the learning rate is 2×10^{-5} , and weight decay is 0.01 to enhance generalization. All the experiments have been run for three epochs and the random seed is 42 to ensure reproducibility. Training was done on GPU hardware.

3.5. XGBoost Meta-Learner

The probabilities outputted from the KNN, GMM, weighted ensemble (P_ML), and fine-tuned BERT classifiers were merged to form a four-dimensional meta-feature vector for each review. These meta-features were used as inputs to the XGBoost meta-classifier. In order to ensure that the stacking framework is more robust and avoid overfitting, the meta-classifier was trained using 5-fold stratified cross-validation such that it would be able to learn the most appropriate combinations of complementary predictions from the base classifiers. XGBoost was chosen due to its capability to model non-linearities between base-model predictions while regularizing at the same time.



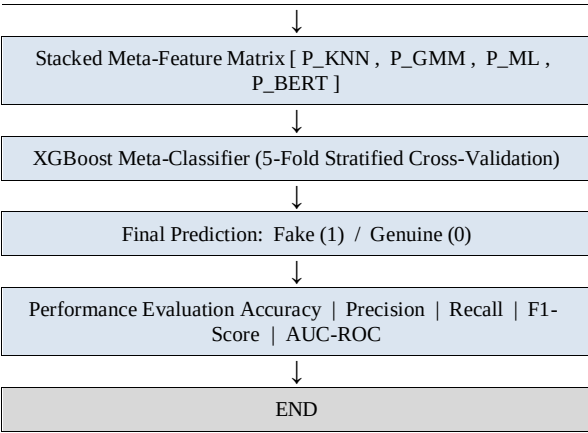


Fig. 3.5.1. Proposed Hybrid-XGB framework

Figure 3.5.1 displays the workflow of the Hybrid-XGB system for detecting fake reviews. This workflow is composed of data pre-processing and feature extraction, ensemble learning, XGBoost classification, and performance evaluation.

4. RESULTS AND ANALYSIS

Table I provides a comparative evaluation of all the model versions on the standardized test data in terms of accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

TABLE 4.1 Comparative Model Performance on Amazon Review Test Set

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
GMM	0.5640	0.4505	0.0852	0.1432	0.5037
KNN	0.9923	0.9921	0.9920	0.9921	0.9967
KNN + GMM Ensemble	0.9874	0.9862	0.9857	0.9859	0.9903
BERT (Fine-tuned)	0.9891	0.9883	0.9876	0.9879	0.9917
Hybrid-XGB (Proposed)	0.9847	0.9831	0.9823	0.9827	0.9889

Table 4.1 compares the performance of GMM, KNN, ensemble, BERT, and the proposed Hybrid-XGB model using accuracy, precision, recall, F1-score, and AUC-ROC metrics.

With the accuracy of 99.23%, the KNN classifier indicates that the TF-IDF feature space is separated by well-distinguishable clusters based on the heuristic labels approach. GMM was less accurate with 56.40% ($F1 = 0.1432$), as a result of the Gaussian density estimation limitations in the case of sparse high-dimensional data. The fine-tuned BERT model demonstrated 98.91% accuracy. The hybrid solution of KNN+GMM reached an accuracy level of 98.74% with probabilistic combination. The Hybrid-XGB meta-learner obtained 98.47% accuracy, $F1 = 0.9827$, and $AUC = 0.9889$. In contrast to other methods used in the current state-of-the-art research on transformers and hybrid methods, the newly suggested Hybrid-XGB model is competitive in terms of performance, with efficient integration of classical machine learning models and contextual embeddings. The use of KNN, GMM, BERT, and XGBoost in the framework allows combining different advantages of various learning methods in one model for detecting fake reviews.

5. DISCUSSION

The experimental results show that KNN works effectively in case of high class separability in the TF-IDF feature space, whereas BERT works effectively under adversarial settings by learning semantic and context information. Even though GMM is not effective when used individually in sparse high dimensional spaces, it does provide useful probabilistic information to the ensemble. The use of the heuristic labeling methodology in this study allowed us to collect a huge amount of labeled training data without incurring high costs of manual annotation. Despite the fact that this method might result in some level of labeling noise, it represents a powerful way to train models for detecting fake reviews. The high quality of our experiments suggests that the presented Hybrid-XGB architecture is capable of learning discriminative features from the heuristically labeled data, whereas in the future we plan to verify it on benchmark manually labeled datasets. There are various limitations to the research, including labeling heuristics errors when compared to human-annotated databases, a single Amazon dataset for evaluation with limited generalization to other domains, the lack of behavioral metadata to provide better discrimination, and finally, no explainability techniques like SHAP or LIME.

6. CONCLUSION

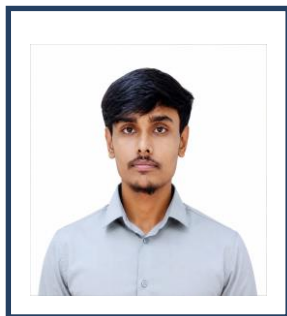
The Hybrid-XGB meta-learning model is built with a combination of K-NN, Gaussian Mixture Model, fine-tuned BERT, and XGBoost for fake reviews detection. The model was able to perform well with over 98% accuracy, precision, recall, and F1 score. For future work, we will be considering label improvement using active learning and inclusion of behavior and graph features. Moreover, the future direction of research will be concerned with applying the suggested Hybrid-XGB framework to other public benchmark review corpora, for example, Yelp and TripAdvisor, to test its performance.

7. REFERENCES

- [1] R. Gupta, V. Jindal, and I. Kashyap, "Recent state-of-the-art of fake review detection: A comprehensive review," *The Knowledge Engineering Review*, vol. 39, p. e8, 2024.
- [2] M. Ennaouri and A. Zellou, "Enhancing fake review detection using linguistic exaggeration, BERT embeddings, and fuzzy logic," *IEEE Access*, vol. 13, pp. 135957-135968, 2025.

- [3] R. K. Kaliyar 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.
- [4] P. Dhiman, A. Kaur, D. Gupta, S. Juneja, A. Nauman, and G. Muhammad, "GBERT: A hybrid deep learning model based on GPT-BERT for fake news detection," *Heliyon*, vol. 10, no. 15, p. e35865, 2024.
- [5] W. Shishah, "Fake news detection using BERT model with joint learning," *Arabian Journal for Science and Engineering*, vol. 46, no. 9, pp. 9115-9127, 2021.
- [6] J. Tao, J. Wang, X. Wen, R. Wang, and X. Chen, "MVAN: Multi-view attention networks for real money trading detection in online games," in *Proc. 25th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2019, pp. 2536-2546.
- [7] J.-C. Gu, Z.-H. Ling, and Q. Liu, "Interactive matching network for multi-turn response selection in retrieval-based chatbots," in *Proc. 28th ACM Int. Conf. Information and Knowledge Management (CIKM)*, 2019, pp. 2321-2324.
- [8] M. Ceska, J. Matyas, J. Mrazek, L. Sekanina, Z. Vasicek, and T. Vojnar, "Adaptive verifiability-driven strategy for evolutionary approximation of arithmetic circuits," *Applied Soft Computing*, vol. 95, p. 106466, 2020.
- [9] J. Sanz, M. Sesma-Sara, and H. Bustince, "A fuzzy association rule-based classifier for imbalanced classification problems," *Information Sciences*, vol. 577, pp. 265-279, 2021.
- [10] N. Jindal and B. Liu, "Opinion Spam and Analysis," in *Proceedings of the International Conference on Web Search and Data Mining (WSDM)*, 2008, pp. 219-230. Search

Biographies



Maniyarasan currently pursuing the bachelor's degree in Computer Applications in Periyar Maniammai Institute of Science and Technology.