JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

AgriMisinfo-ID: A Multi-Platform Dataset and Ensemble Transformer-Based Detection System for Agricultural Misinformation in Indonesian Social Media

Brian Rizqi Paradisiaca Darnoto [1], Nelly Oktavia Adiwijaya [1], Dony Bahtera Firmawan [1], Fadhel Akhmad Hizham [1,*], Nandini Putri Hanifa Jannah [1], Talitha Puspitasari [2], and Fabyan Yastika Permana [2]

[1] Department of Informatics, University of Jember, Jember 68121, Indonesia;
e-mail : brianrizqi@unej.ac.id; nelly.oa@unej.ac.id; donybf@unej.ac.id; fadhel.ilkom@unej.ac.id; 232410103024@mail.unej.ac.id

[2] Department of Information System, University of Jember, Jember 68121, Indonesia;
e-mail : 242410101017@mail.unej.ac.id; 242410101041@mail.unej.ac.id

***** Corresponding Author : Fadhel Akhmad Hizham

Abstract: Agricultural misinformation on social media poses risks to Indonesian food security and farmer livelihoods. False claims about fertilizers, pesticides, crop varieties, and farming practices can spread rapidly across social media platforms such as Instagram, YouTube, and Facebook, as well as through publicly available datasets, potentially influencing agricultural decisions and outcomes. This paper introduces AgriMisinfo-ID, a bilingual, multi-platform dataset for agricultural misinformation detection, containing 5,288 labeled samples collected from social media and supplementary public datasets across Indonesian agricultural contexts. A hybrid detection system is proposed that combines an ensemble of fine-tuned Transformer models, IndoBERT and XLM-RoBERTa, with a Knowledge Verification module that cross-references agricultural claims against Wikidata and Wikipedia. Training uses Focal Loss to address class imbalance, together with GPT-4o-mini-based paraphrase augmentation for minority classes. Across three random seeds, the weighted ensemble achieves an F1-Macro of 0.5870 ± 0.0028 and an accuracy of 0.8027 ± 0.0039 on the test set, outperforming the individual models, a TF-IDF/SVM baseline, and an equal-weight ensemble in terms of F1-Macro. The Knowledge Verification module provides evidence-based verdicts that can support the inspection and auditability of model decisions. This work provides a reproducible benchmark for agricultural misinformation research in bilingual, low-resource settings.

Keywords: Agricultural misinformation detection; Ensemble learning; IndoBERT; Indonesian NLP; Knowledge verification; Misinformation detection; Social media; XLM-RoBERTa.

Received: July, 1st 2026
Revised: August, 2nd 2026
Accepted: August, 8th 2026
Published: August, 28th 2026

1. Introduction

False information can spread faster than corrections, and farming communities may bear tangible economic consequences as a result. A viral post claiming that urea fertilizer permanently poisons soil, or that organic pesticides are always safer than synthetic ones, can influence purchasing decisions across farming networks within a short period. By the time a correction circulates, farmers may already have switched products, delayed planting, or abandoned pest-management practices that were effective [1]. Indonesia makes this problem particularly relevant. More than 33 million farmers work in a sector that contributes roughly 13% of GDP, giving agricultural misinformation potentially significant economic consequences. Indonesia also has 212 million internet users and more than 167 million active social media accounts [2]. Facebook and Instagram reach over 80% of Indonesian internet users, while YouTube ranks second in national platform visits. Farmers use these channels to seek information about fertilizer dosages, crop disease symptoms, and pest-control practices, often without reliable means of verifying the information they encounter.

Automated misinformation detection has advanced rapidly in recent years [3], but much of this progress has focused on political news and general hoax detection. In the Indonesian context, previous work has addressed political hoaxes [4] and general news classification [5], while agricultural claims remain largely unexplored. Agricultural misinformation presents distinct challenges: it involves domain-specific vocabulary that may not be adequately captured by general-purpose models, can combine true and false elements within a single claim, and may require a five-category label space that binary hoax classifiers cannot adequately represent [6]. Most Indonesian misinformation detection systems use binary labels, such as true or false, which may be too coarse for agricultural content. A post claiming that urea is safe at twice the recommended dose, for example, may contain a factual element while leading to a potentially harmful conclusion. Agricultural misinformation therefore benefits from a finer-grained classification scheme that distinguishes misleading, unverified, and satirical content, together with models capable of differentiating among these categories.

Building such a system is increasingly practical. Pre-trained Transformer models, particularly IndoBERT [7] and XLM-RoBERTa [8], provide strong Indonesian and multilingual text representations, respectively. Structured knowledge sources such as Wikidata and Wikipedia can provide evidence for verifying agricultural claims. Evidence retrieval has also been identified as a practical approach for improving the interpretability of misinformation detection systems [9]. Focal Loss [10] can help address severe class imbalance without requiring extremely large amounts of labeled data. These components are therefore integrated into a unified detection pipeline for agricultural misinformation. The main contributions of this paper are as follows:

- A domain-specific dataset is introduced, consisting of 5,288 labeled samples across five categories collected from Instagram, YouTube, Facebook, and supplementary public sources within Indonesian agricultural contexts.
- An ensemble-based detection approach is developed by combining IndoBERT and XLM-RoBERTa with Focal Loss and GPT-4o-mini-based paraphrase augmentation to address multilingual representation and class imbalance.
- A Knowledge Verification module is incorporated to query Wikidata and Wikipedia and provide evidence-based verdicts and human-readable explanations.

The remainder of this paper is organized as follows. Section 2 reviews related work on misinformation detection, Indonesian-language misinformation detection, knowledge graph-enhanced fact verification, and data augmentation for imbalanced NLP. Section 3 presents the proposed methodology, including dataset construction, data augmentation, the ensemble Transformer architecture, and the Knowledge Verification module. Section 4 presents and discusses the experimental results, including overall performance, per-class analysis, confusion matrices, ablation experiments, and Knowledge Verification analysis. Section 5 discusses the limitations and practical implications of the study. Finally, Section 6 concludes the paper and outlines directions for future research.

2. Literature Review

Misinformation detection has attracted substantial research across computational linguistics, information systems, and related fields. This section reviews four bodies of work most relevant to the proposed approach: general misinformation and fake news detection using Transformer models, Indonesian-language misinformation detection, knowledge graph-enhanced fact verification, and data augmentation for imbalanced NLP.

2.1. Misinformation and Fake News Detection

Early fake news detection approaches relied on handcrafted features combined with classifiers such as SVM and logistic regression [11]. Shu et al. [1] organized this body of work into four main families: knowledge-based, style-based, propagation-based, and credibility-based methods, each based on different assumptions about the signals that distinguish false from true content. The introduction of BERT [12] substantially advanced this field; Kaliyar et al. [13] showed that BERT-based approaches outperformed earlier methods on fake news classification, while subsequent studies explored further improvements using different Transformer-based configurations. Dhiman et al. [14] proposed GBERT, a hybrid approach that combines GPT-based generation with BERT classification for richer feature extraction.

Aljawarneh and Swedat [15] demonstrated that progressive fine-tuning of BERT can improve performance on social media datasets. Essa et al. [16] combined BERT with LightGBM and reported consistent gains compared with either component alone. Ilie et al. [17] benchmarked deep learning architectures across multiple context-aware misinformation datasets, providing useful reference points for model selection.

Ensemble approaches have also become increasingly common in fake news detection. Lin et al. [18] showed that combining BERT-based ensembles with sentiment analysis can improve the identification of harmful news content. Al-Alshaqi et al. [19] demonstrated that combining multiple Transformer models with traditional machine learning classifiers can provide reliable performance across diverse modalities. Almandouh et al. [20] evaluated ensemble approaches across multiple benchmark configurations, while Khan et al. [21] extended ensemble-based approaches to hybrid Transformer architectures for social media misinformation detection. Kumar et al. [22] combined BERT and RoBERTa with graph neural networks and reported strong binary classification results on FakeNewsNet. Mridha et al. [23] reviewed deep learning approaches for fake news detection, while Plikynas et al. [24] found that misleading content can be more difficult to detect than outright false claims, highlighting the importance of fine-grained classification.

Knowledge-enhanced detection provides another direction for addressing limitations of purely neural approaches, particularly when additional evidence is needed to support model predictions. Seddari et al. [25] showed that combining linguistic analysis with knowledge-base lookups can improve both accuracy and interpretability on social media datasets. Xie et al. [26] incorporated knowledge graphs into heterogeneous graph neural networks for fake news detection and reported improvements in recall for minority categories. Joshi et al. [27] developed a multi-platform explainable detection system using domain adaptation and LIME, with user studies reporting increased trust when evidence citations were provided. Kuntur et al. [28] surveyed the use of large language models for both misinformation detection and generation, highlighting the rapidly evolving nature of the field. Gongane et al. [9] reviewed explainability techniques for misinformation detection and identified evidence retrieval as one of the practical approaches for supporting interpretable detection systems.

2.2. Indonesian Language Misinformation Detection

Indonesian misinformation detection has grown considerably, although existing coverage remains relatively narrow. Rahmawati et al. [4] fine-tuned IndoBERT for political hoax detection, establishing IndoBERT as an important baseline for Indonesian text classification. Sinapoy et al. [29] compared LSTM and IndoBERT for Twitter hoax detection and reported 92.07% accuracy for binary classification, although that task involved only two categories. Nayoga et al. [5] evaluated CNN, LSTM, and Transformer configurations on an Indonesian news hoax corpus. Praha et al. [30] investigated CNN-LSTM transfer learning for fake news classification and reported stronger performance for LSTM-based configurations. Safira and Setiawan [31] compared Bi-LSTM and 1D-CNN approaches for social media hoax detection, while Pekandi [32] evaluated IndoBERT against ensemble and CNN-LSTM variants and reported 92.24% accuracy on 6,120 articles. Sagama and Alamsyah [33] applied BERT and RoBERTa to Indonesian online toxicity classification, further demonstrating the applicability of Transformer-based models to Indonesian text.

Two additional studies extend Indonesian NLP classification to directions relevant to this work. Sirusstara et al. [34] applied RoBERTa to clickbait detection in Indonesian news headlines, demonstrating the applicability of Transformer models to subtler forms of content manipulation. Laimeheriwa et al. [35] surveyed Indonesian hoax analysis and fake news detection using deep learning and identified challenges including limited labeled data and domain-specific vocabulary gaps. Taken together, these studies demonstrate the applicability of Transformer-based models to Indonesian text classification while also highlighting the continued need for domain-specific datasets and methods. However, the reviewed studies do not specifically address agricultural misinformation using a five-class label scheme together with structured knowledge verification. AgriMisinfo-ID addresses these gaps by combining a domain-specific agricultural misinformation dataset, fine-grained classification, and knowledge verification.

Firmawan [36] demonstrated that SHAP can produce higher faithfulness scores than attention-based explanations across Transformer-based NLP models under domain shift, providing empirical support for the use of perturbation-based evidence scoring in the

Knowledge Verification module. Darnoto [37] applied transfer fine-tuning guided by linguistic similarity to Indonesian regional-language sentiment analysis and showed that IndoBERT and XLM-R provide complementary strengths under low-resource conditions, providing supporting evidence for the use of complementary Transformer models in the proposed ensemble.

2.3. Knowledge Graph-Enhanced Fact Verification

Fact verification using structured knowledge sources predates the widespread adoption of deep learning. FEVER [38] established Wikipedia-grounded verification as a standard evaluation framework and remains widely used for benchmarking fact-checking systems. Kim et al. [39] extended this direction to multi-hop reasoning over knowledge graphs, showing how structured sources can support the verification of complex claims. Chen et al. [40] reviewed knowledge graph completion methods that underpin evidence retrieval in automated fact-checking pipelines. Xie et al. [26] applied knowledge graph enhancement specifically to fake news detection and reported improved recall for minority categories, which is particularly relevant to the class imbalance encountered in AgriMisinfo-ID.

The motivation for combining neural prediction with knowledge verification lies not only in classification performance but also in the availability of supporting evidence. Seddari et al. [25] reported measurable accuracy improvements from combining linguistic analysis with knowledge-base lookups while also highlighting the interpretability benefits of the hybrid approach. Joshi et al. [27] found that explainable systems providing cited evidence can increase user trust compared with black-box classifiers across multiple social media platforms. Gongane et al. [9] reviewed explainability techniques and identified evidence retrieval as one of the practical approaches for deployment-oriented misinformation detection. Golda et al. [41] further discussed the potential of hybrid systems that combine neural models with structured knowledge, particularly in high-stakes domains. These studies provide supporting motivation for incorporating structured knowledge verification alongside neural misinformation classification.

2.4. Data Augmentation for Imbalanced NLP

Class imbalance is a persistent challenge in misinformation datasets, where false and satirical content may occur much less frequently than true or unverified posts. Sahin [42] benchmarked text augmentation techniques for low-resource NLP across multiple languages and found that augmentation can improve F1 performance on minority classes, with the benefits depending on the severity of class imbalance. Muftie and Haris [43] evaluated IndoBERT-based selective word insertion for Indonesian text classification and reported improvements over random-insertion baselines on Twitter data. Kim et al. [39] showed that augmented training combined with focused contrastive loss can achieve strong performance on imbalanced classification tasks, providing supporting motivation for combining augmentation with Focal Loss in AgriMisinfo-ID.

Among different augmentation strategies, LLM-based paraphrase generation provides an effective way to expand minority classes while preserving the semantic content of the original samples. Compared with rule-based approaches such as synonym substitution or random insertion, LLM-generated paraphrases can introduce lexical variation while retaining domain-specific terminology and sentence structure. This consideration is particularly relevant to agricultural text, where terms such as urea, NPK, and varietas unggul need to be preserved accurately to avoid introducing noise into the training data.

A critical consideration in text augmentation is the potential introduction of label noise. Paraphrases that alter the factual meaning of a claim may effectively change its original label and consequently corrupt the training signal. In this study, GPT-4o-mini was instructed to preserve the original claim's meaning and agricultural context while varying its wording. A quality-filtering procedure was then applied to exclude paraphrases whose semantic content diverged substantially from the source according to a sentence-level semantic similarity threshold. Only paraphrases that passed this filtering process were included in the training set.

3. Method

This section describes the four main components of AgriMisinfo-ID: dataset construction, data augmentation, the ensemble Transformer architecture, and the Knowledge Verification module.

3.1. Dataset Construction

AgriMisinfo-ID was constructed through a five-stage pipeline comprising data collection, preprocessing, annotation, dataset splitting, and training-set augmentation. Figure 1 illustrates the overall dataset construction pipeline and the flow of posts through these stages. Data were collected from social media sources and supplementary public datasets, followed by preprocessing and annotation. After annotation, the dataset was divided into training, validation, and test partitions using stratified sampling. Data augmentation was then applied exclusively to the training partition to address class imbalance.

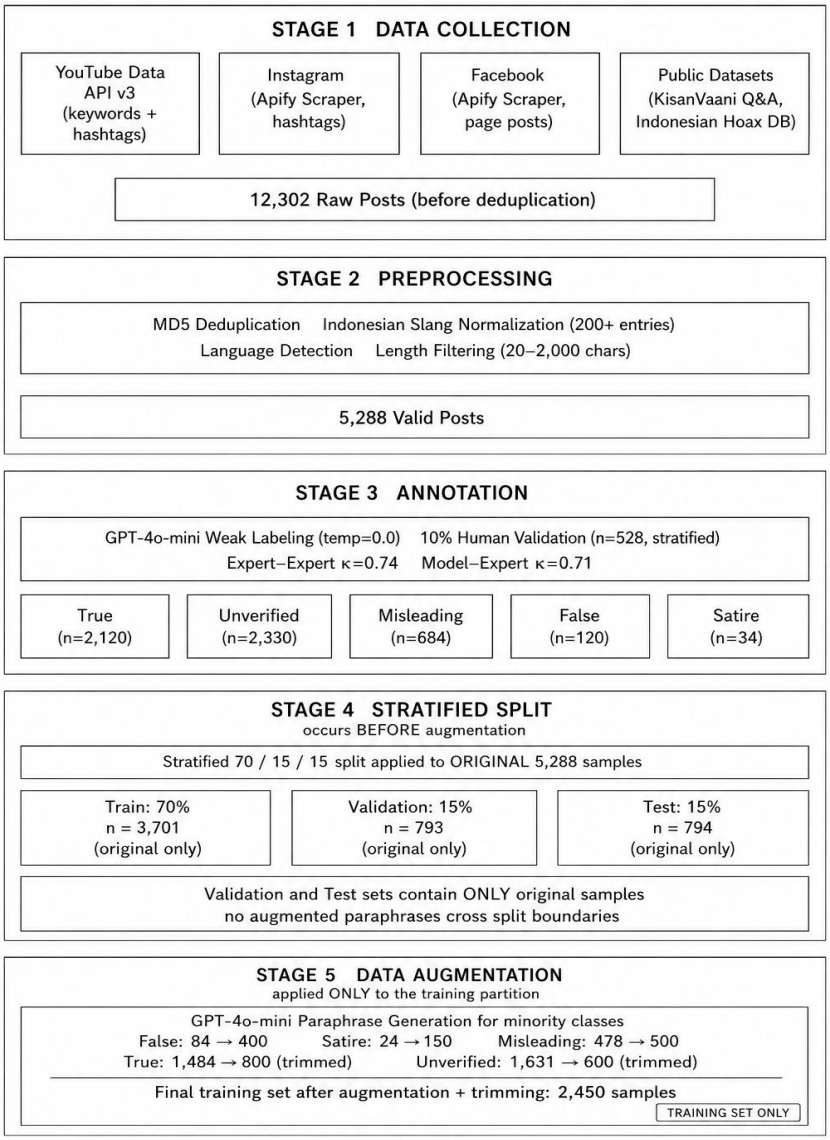

Figure 1. AgriMisinfo-ID dataset construction pipeline across five stages.

3.1.1. Data Collection

Agricultural content was actively collected between January 2020 and December 2024 from four sources: YouTube through Data API v3, Instagram and Facebook through Apify cloud scraping, and supplementary public datasets including KisanVaani English Q&A and Indonesian hoax collections. The collection used 36 domain-specific keywords and hashtags

in Indonesian and English, covering fertilizers (pupuk, urea, NPK, kompos), pesticides (pestisida, insektisida, fungisida), crop diseases (hama, penyakit tanaman), and farming practices (pertanian organik, hidroponik, varietas unggul). The raw collection produced 12,302 posts before deduplication. To support reproducibility, the data collection scripts, preprocessing pipeline, and annotation guidelines are publicly available through the project repository, together with the dataset released under a CC BY-SA 4.0 license.

Figure 1 shows how the collected posts move through the preprocessing stages. Preprocessing included MD5-based deduplication, Indonesian slang normalization using a 200-entry dictionary, language detection, and length filtering to retain posts between 20 and 2,000 characters. After all filtering steps were applied, 5,288 posts remained.

Keyword selection followed a two-stage process. In the first stage, an initial keyword set was derived from Indonesian Ministry of Agriculture extension materials and three previous studies on agricultural misinformation in South and Southeast Asia. In the second stage, the keyword set was expanded by analyzing frequently co-occurring terms in a 500-post pilot collection. Terms that frequently appeared in agricultural misinformation contexts but were absent from the initial list were added. The final set of 36 keywords covers major categories of agricultural misinformation, including fertilizer safety, pesticide toxicity, GMO crop claims, crop disease remedies, and farming-practice myths.

3.1.2. Annotation

GPT-4o-mini, accessed through the OpenRouter API, was used for weak labeling and was prompted to act as an agricultural fact-checking expert. The use of LLMs as annotators has been investigated in previous studies. Gilardi et al. [44] showed that ChatGPT can outperform crowd workers on annotation tasks including stance, topic, and frame detection across 6,183 tweets and news articles, while He et al. [45] demonstrated that GPT-3.5 can match or exceed crowdsourced annotators through an explain-then-annotate framework.

Five categories were assigned: True, referring to factually accurate claims; False, referring to claims that contradict established agricultural science; Misleading, referring to partially correct claims framed deceptively; Unverified, referring to claims that cannot be confirmed or refuted using the available sources; and Satire, referring to clearly humorous content that is not intended as factual information. Two domain experts independently reviewed 528 samples selected through stratified random sampling by class: True (212), Unverified (232), Misleading (68), False (12), and Satire (4). Cohen's Kappa between the two experts was 0.74, indicating reliable inter-annotator agreement, while the model–expert Kappa between GPT-4o-mini and the expert consensus was 0.71, indicating substantial agreement [46].

The five label categories were defined with reference to the agri-food misinformation taxonomy. True posts contain claims that can be verified against authoritative agricultural sources. False posts directly contradict established agricultural science. Misleading posts contain accurate facts framed to lead toward an incorrect conclusion. For example, a post may state that a fertilizer increases yield by 300% under controlled conditions while implying that the same increase can be achieved under typical smallholder conditions. Unverified posts cannot be confirmed or refuted using the available sources. Satire posts are clearly humorous and are not intended as factual claims.

Annotator training was conducted before the validation review. Each expert received the category definitions, five annotated examples for each category, and a borderline-case guide covering common ambiguities, particularly between Misleading and Unverified. Disagreements occurred on 73 of the 528 validated samples (13.8%) and were resolved through a third adjudication round in which both experts discussed each case until consensus was reached.

3.1.3. Label Distribution

The final dataset contains 5,288 samples distributed across five categories: True (40.1%, n = 2,120), Unverified (44.0%, n = 2,330), Misleading (12.9%, n = 684), False (2.3%, n = 120), and Satire (0.6%, n = 34). Posts span 2014 to 2024 by their original creation date. The posts from 2014–2019 originate from supplementary public datasets, including KisanVaani Q&A and Indonesian hoax collections, whereas scraped content covers 2020–2024. Indonesian-language posts account for 37.1% of the dataset, while English-language posts account for 62.9%. Accordingly, AgriMisinfo-ID is treated as a bilingual dataset covering Indonesian agricultural contexts.

The raw source distribution comprises 4,098 YouTube posts, 608 Instagram posts, 5 Facebook posts, and 1,221 posts from supplementary public datasets. Because some posts appeared in both scraped and public dataset collections, the source counts are not mutually exclusive: the raw counts sum to 5,932 (4,098 + 608 + 5 + 1,221), and 644 posts that appeared in multiple source categories were identified and removed by MD5-based deduplication, yielding 5,288 unique posts. The dataset was subsequently divided using stratified sampling into training (70%, n = 3,701), validation (15%, n = 793), and test (15%, n = 794) partitions.

The relatively high proportion of Unverified samples (44.0%) reflects the difficulty of verifying agricultural claims in a low-resource setting. Many agricultural posts contain plausible claims about crop varieties, soil treatments, or pest-control methods that cannot be confirmed or refuted without access to localized agricultural extension information, which is not always available in structured form. This distribution therefore highlights the verification challenge addressed by the Knowledge Verification module.

The False (2.3%) and Satire (0.6%) categories are substantially underrepresented relative to the other categories. This class distribution creates a potential bias toward majority classes during model training. The data augmentation strategy described in Section 3.2 and the Focal Loss objective described in Section 3.3.1 are therefore used to mitigate the effects of class imbalance.

3.2. Data Augmentation

The False and Satire categories contained only 84 and 24 training samples, respectively, creating a substantial imbalance relative to the majority classes. GPT-4o-mini was used to generate three to five paraphrases per minority-class sample while preserving the agricultural context and varying the wording. Figure 2 presents the class distribution before and after augmentation on the training split.

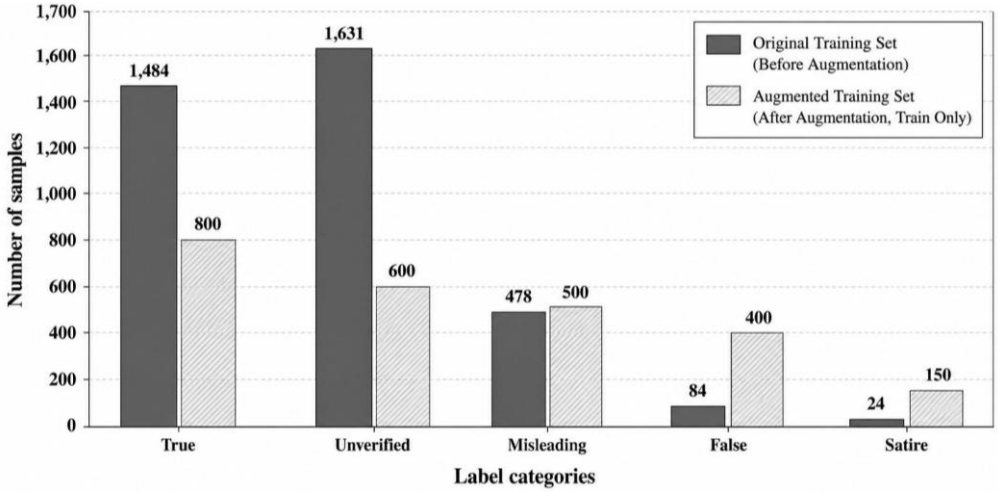

Figure 2. Training-set class distribution before and after data augmentation.

Augmentation was applied exclusively to the training partition after stratified splitting. The validation and test sets contain only original posts, and no augmented paraphrase was allowed to cross partition boundaries. False samples increased from 84 to 400, Satire from 24 to 150, and Misleading from 478 to 500. True and Unverified samples were trimmed to 800 and 600, respectively, to reduce majority-class dominance. After quality filtering, 464 augmented samples were retained (False: +316, Satire: +126, Misleading: +22) and added to the training data. Of the 3,701 original training samples, 1,715 majority-class samples were removed during trimming (True: 1,484 → 800, removing 684; Unverified: 1,631 → 600, removing 1,031). The final training set therefore contains 1,986 original retained samples plus 464 augmented paraphrases, for a total of 2,450 samples (True 800 + Unverified 600 + Misleading 500 + False 400 + Satire 150 = 2,450).

Quality control was performed in two stages. First, GPT-4o-mini was prompted to preserve the original claim meaning, agricultural context, and label while varying only the wording and sentence structure. Second, a semantic similarity filter was applied. Paraphrases with

cosine similarity below 0.65 or above 0.98 relative to the source sentence, computed using multilingual sentence embeddings, were discarded. The lower threshold was used to ensure sufficient semantic similarity, while the upper threshold was intended to exclude near-exact duplicates that provide limited additional training value. After filtering, 464 paraphrases passed the quality-control procedure.

The target class distribution after augmentation and trimming was selected to balance minority-class exposure while preserving sufficient majority-class training data. False and Satire were increased to 400 and 150 samples, respectively, while True and Unverified were reduced to 800 and 600 samples. These target values were selected empirically by comparing validation F1-Macro across several configurations before selecting the final configuration.

3.3. System Architecture

The overall architecture of the proposed AgriMisinfo-ID detection system is illustrated in Figure 3. The system consists of two parallel branches: an ensemble Transformer branch for text classification and a Knowledge Verification branch for evidence-based claim verification. The outputs of the two branches are subsequently combined through a fusion layer to produce the final prediction, confidence score, and explanation.

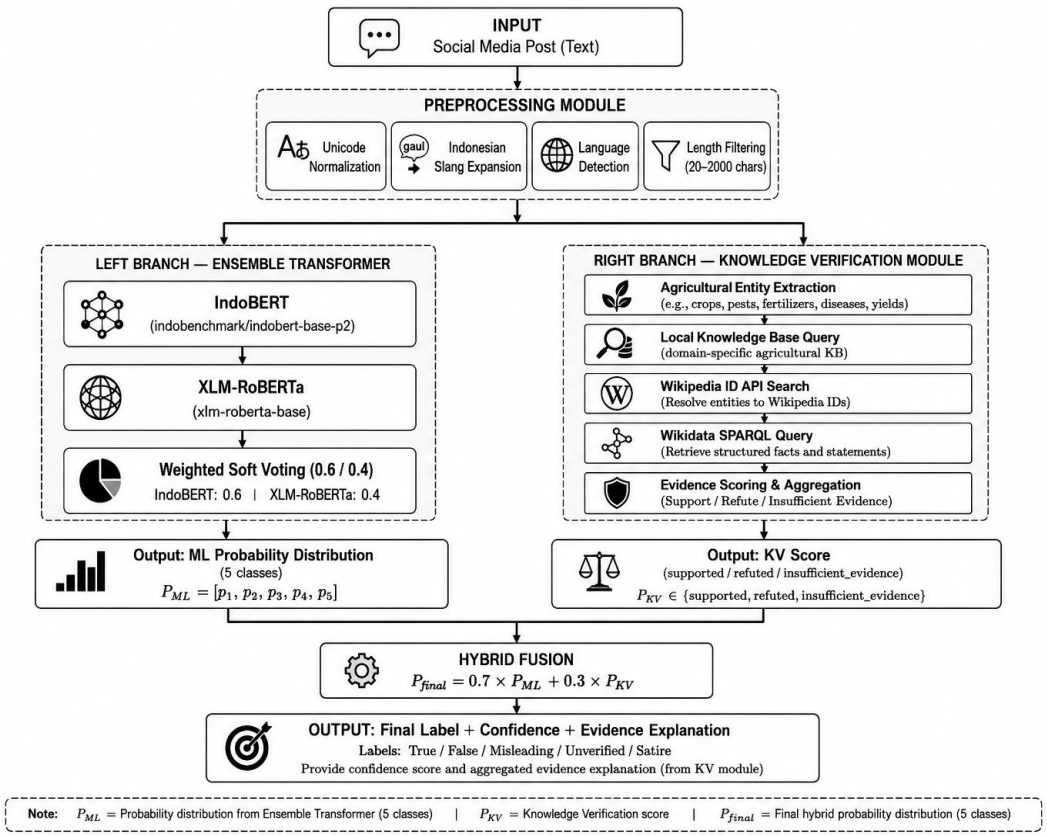

Figure 3. Overall architecture of the AgriMisinfo-ID detection system.

Each component addresses a specific characteristic of the dataset and classification task. IndoBERT is used because it was pre-trained on Indonesian Wikipedia, news articles, and social media text, providing representations suited to Indonesian linguistic patterns, including informal language, abbreviations, and code-switching. XLM-RoBERTa is included because 62.9% of the dataset is in English, providing multilingual representations that complement the Indonesian-focused capabilities of IndoBERT. The two models are therefore used as complementary components of the ensemble.

Focal Loss is used instead of standard cross-entropy to mitigate the effects of severe class imbalance. Standard cross-entropy may allow majority-class examples to dominate the optimization process, whereas Focal Loss down-weights easy examples and places greater emphasis on difficult cases, including minority-class samples. Knowledge Verification is

incorporated because agricultural claims may depend on specific numerical facts or biological relationships that cannot necessarily be verified from textual patterns alone.

3.3.1. Ensemble Transformer

IndoBERT [7] was pre-trained on Indonesian Wikipedia, news, and social media text, while XLM-RoBERTa [8] was trained across 100 languages using 2.5 terabytes of CommonCrawl data. Both models were fine-tuned for five-class classification using a maximum sequence length of 256, an effective batch size of 64, consisting of a per-device batch size of 32 with two gradient-accumulation steps, a cosine learning-rate schedule with 10% warmup, weight decay of 0.01, label smoothing of 0.1, and a maximum of 10 epochs with early stopping based on validation F1-Macro with a patience of three epochs. Training was performed on an NVIDIA A100-SXM4-40GB GPU.

Focal Loss [9] was configured with a gamma value of 2.0 and class-balanced alpha weights. The loss down-weights easy examples and directs greater gradient emphasis toward difficult minority-class cases [47]. The ensemble combines the two model probability distributions using weighted soft voting $P_{\text{ensemble}} = 0.6P_{\text{IndoBERT}} + 0.4P_{\text{XML-R}}$. The 0.6/0.4 weighting was selected through a grid search on the validation set with increments of 0.1. IndoBERT received the higher weight because it achieved stronger F1-Macro performance on the validation data, particularly for the Misleading and Unverified categories. This weighting is also consistent with findings reported by Darnoto [37] regarding the complementary strengths of IndoBERT and XLM-R in Indonesian text classification.

Label smoothing with a smoothing rate of 0.1 was applied to reduce overconfident predictions during training. Gradient accumulation over two steps was used to obtain an effective batch size of 64 without exceeding GPU memory limitations. Early stopping with a patience of three epochs based on validation F1-Macro was used to limit overfitting while allowing sufficient training time for minority-class representations to stabilize. The final ensemble weights were selected exclusively using the validation set. The grid search evaluated combinations from 0.1/0.9 to 0.9/0.1 and selected the configuration that maximized validation F1-Macro.

3.3.2. Knowledge Verification Module

The Knowledge Verification module performs evidence retrieval to complement the predictions generated by the Transformer ensemble. Entity extraction uses pattern matching against a 50-term agricultural dictionary covering fertilizers, pests, crops, and agrochemicals. The dictionary is supported by a local knowledge base containing 15 entity profiles derived from Indonesian agricultural extension publications. Matched entities trigger searches across three knowledge sources: the local knowledge base, Indonesian Wikipedia through the MediaWiki API, and Wikidata through SPARQL [48].

Retrieved evidence items are scored according to keyword overlap between the input claim and the retrieved snippets. A knowledge score (KS) above 0.65 produces a supported verdict, a score below 0.35 produces a refuted verdict, and scores between these thresholds produce an insufficient_evidence verdict. Each verdict is accompanied by a cited explanation [9]. The Knowledge Verification output is then combined with the machine-learning prediction using $P_{\text{final}} = 0.7P_{\text{ML}} + 0.3P_{\text{KV}}$. When no agricultural entities are detected, the system falls back to the ML-only prediction.

The three knowledge sources serve complementary purposes. The local knowledge base provides high-precision evidence through 15 manually curated entity profiles derived from Indonesian Ministry of Agriculture publications. These profiles cover commonly misrepresented agricultural facts, including fertilizer composition, recommended pesticide dosages, and crop-variety characteristics. Wikipedia provides broader coverage but may return snippets that do not directly address the claim. Wikidata provides structured fact triples that can be queried precisely but is limited to entities represented in the knowledge graph. Evidence from the three sources is aggregated into a single knowledge score, with greater weight assigned to evidence from the local knowledge base.

The 0.7/0.3 ML–KV fusion weights were selected through a validation-set grid search that evaluated combinations from 0.5/0.5 to 0.9/0.1 in increments of 0.05. The selected configuration provided a balance between the contribution of the Knowledge Verification signal and the overall classification performance.

4. Results and Discussion

This section evaluates the proposed AgriMisinfo-ID system across five dimensions: experimental setup, overall model performance, per-class performance, confusion-matrix analysis, ablation of the training components, and Knowledge Verification analysis. Because the dataset is substantially imbalanced, F1-Macro is treated as the primary evaluation metric, while accuracy and F1-Weighted are reported as complementary measures.

4.1. Experimental Setup

All experiments were conducted on an NVIDIA A100-SXM4-40GB GPU using PyTorch 2.6.0 and Hugging Face Transformers 4.35. All reported results are from the primary evaluation run. To assess robustness, the ensemble was additionally evaluated across three random seeds (42, 123, 456), yielding F1-Macro of 0.5870 ± 0.0028 and accuracy of 0.8027 ± 0.0039, confirming stability of the reported results. Model selection and hyperparameter tuning were performed using the validation set, with F1-Macro as the selection criterion. The test set was used only after the model configurations and hyperparameters had been finalized. The baseline classifier used TF-IDF features based on unigrams and bigrams, with a maximum vocabulary of 50,000 features and sublinear term-frequency scaling, followed by a LinearSVC classifier with class-balanced weights.

F1-Macro was selected as the primary metric because of the pronounced class imbalance in the test set. Satire accounts for only 0.6% of the test samples (n = 5), while False accounts for 2.3% (n = 18). Under such a distribution, accuracy can obscure poor performance on minority classes. For example, a classifier that always predicts the majority class, Unverified, would achieve 44.1% accuracy but an F1-Macro of only 0.12. F1-Macro assigns equal importance to all classes regardless of their support and therefore provides a more informative assessment of whether the models can distinguish the five misinformation categories. F1-Weighted is additionally reported because it reflects performance while accounting for the observed class distribution.

The TF-IDF/LinearSVC configuration provides a conventional text-classification reference against which the benefits of contextual Transformer representations can be assessed. Class-balanced weighting was applied to the SVM objective to reduce the dominance of the majority classes during training. The resulting baseline therefore provides a stronger comparison than an unweighted traditional classifier while remaining computationally simpler than the proposed Transformer-based models.

4.2. Overall Performance

Table 1 summarizes the overall performance of the evaluated models on the test set. Because the dataset is highly imbalanced, F1-Macro is considered the primary metric, while accuracy and F1-Weighted are reported as complementary measures. The majority-class baseline achieves an F1-Macro of only 0.1224, illustrating the difficulty of obtaining balanced performance across the five classes by relying primarily on the dominant category.

Table 1. Overall performance on the test set

Model	Accuracy	F1-Macro	F1-Weighted
SVM + TF-IDF (Baseline) [11]	0.7406	0.5206	0.7367
IndoBERT [7]	0.7834	0.5809	0.7822
mBERT (bert-base-multilingual-cased)	0.7355	0.4454	0.7274
XLM-RoBERTa [8]	0.7708	0.4978	0.7702
Ensemble (Ours)	0.8048	0.5871	0.8032

The SVM + TF-IDF baseline achieves an F1-Macro of 0.5206, indicating that conventional lexical representations can capture some discriminative patterns in agricultural misinformation. However, IndoBERT achieves a higher F1-Macro of 0.5809, together with an accuracy of 0.7834 and an F1-Weighted score of 0.7822. This improvement suggests that contextual representations are better suited to capturing the linguistic patterns required for distinguishing the five misinformation categories.

mBERT obtains the lowest F1-Macro among the Transformer-based models, at 0.4454, despite achieving an accuracy of 0.7355. XLM-RoBERTa performs better than mBERT, reaching an F1-Macro of 0.4978, but remains below IndoBERT. These results suggest that multilingual pre-training alone does not necessarily provide an advantage for this task. The stronger performance of IndoBERT may be related to its Indonesian-focused pre-training, which is potentially better aligned with the linguistic characteristics of a substantial portion of the dataset.

The ensemble achieves the highest scores across all three evaluation metrics, with an accuracy of 0.8048, an F1-Macro of 0.5871, and an F1-Weighted score of 0.8032. Multi-seed evaluation across three seeds (42, 123, 456) confirmed result stability (F1-Macro: 0.5870 $\pm$ 0.0028). The improvement over IndoBERT in F1-Macro is relatively modest, indicating that the principal benefit of combining the two Transformer models is not a large increase in aggregate performance but potentially more balanced predictions across the five categories. This observation is particularly relevant for the present task because the minority classes are substantially smaller than the True and Unverified categories.

Taken together, the results in Table 1 indicate that Transformer-based models provide stronger macro-averaged performance than the conventional SVM + TF-IDF baseline, with IndoBERT providing the strongest individual-model performance and the ensemble achieving the highest overall F1-Macro. However, the differences among the models should be interpreted alongside the per-class results, as macro-averaged performance alone does not reveal which misinformation categories remain difficult to classify. These class-level differences are examined in Section 4.3.

4.3. Per-Class Performance

Overall F1-Macro provides a useful summary of model performance, but it can conceal substantial differences across individual misinformation categories. Table 2 presents the per-class F1-scores together with the corresponding test-set support, allowing the performance of each category to be interpreted in relation to the number of available test samples.

Table 2. Per-class F1-score on the test set

Label	Support	SVM	IndoBERT	XLM-R	Ensemble
True	318	0.7889	0.8380	0.8308	**0.8549**
False	18	0.1600	**0.3529**	0.2632	0.3125
Misleading	103	0.5370	0.6190	0.5899	**0.6479**
Unverified	350	0.7836	0.8086	0.8053	**0.8345**
Satire	5	0.3333	0.2857	0.0000	0.2857

The ensemble achieves its strongest performance on the two most prevalent categories, True and Unverified, with F1-scores of 0.8549 and 0.8345, respectively. These categories have substantially larger test-set support (n = 318 and n = 350), providing considerably more examples for learning and evaluation. The ensemble also performs best on the Misleading category, achieving an F1-score of 0.6479 compared with 0.6190 for IndoBERT and 0.5899 for XLM-RoBERTa. The 2.89 percentage-point improvement over IndoBERT suggests that combining complementary Transformer representations may be particularly useful for content that contains both factual and misleading elements.

The Misleading category is especially important for this application because such posts may contain genuine agricultural facts while framing them in a way that leads to an incorrect or unsupported conclusion. The relatively lower F1-score compared with True and Unverified therefore indicates that distinguishing subtle semantic manipulation remains challenging even for the ensemble. This result is consistent with the motivation for using a fine-grained five-class formulation rather than a binary true/false classification.

Performance on the False and Satire categories is considerably lower. The ensemble achieves F1-scores of 0.3125 for False and 0.2857 for Satire. However, these values should be interpreted cautiously because the corresponding test supports are only 18 and 5 samples, respectively. The limited number of original minority-class examples means that a small number of prediction changes can substantially affect the resulting F1-score. Therefore, these

results provide evidence of the difficulty of minority-class recognition but should not be interpreted as definitive estimates of generalization performance for these categories.

The comparison among models also reveals different strengths. IndoBERT achieves the highest F1-score on the False category (0.3529), while the ensemble performs better on True, Misleading, and Unverified. XLM-RoBERTa obtains an F1-score of 0.0000 on Satire, whereas IndoBERT and the ensemble both obtain 0.2857. The latter result suggests that the Indonesian-focused representation provided by IndoBERT may offer some advantage for this category, although the extremely small support of five samples prevents a strong conclusion about the underlying cause. Overall, the per-class results show that the ensemble's improvement is concentrated primarily in the better-represented and practically important categories, particularly Misleading. At the same time, the persistent weakness on False and Satire indicates that class balancing and ensemble learning alone do not fully resolve the scarcity of minority-class evidence. This pattern motivates the subsequent confusion-matrix analysis, which examines the specific error patterns underlying these aggregate per-class scores.

4.4. Confusion Matrix Analysis

Figures 4–6 present the normalized confusion matrices for IndoBERT, XLM-RoBERTa, and the proposed ensemble, respectively. All percentages shown in the confusion matrices are row-normalized, with each value representing the proportion of all samples belonging to the corresponding true class. Thus, the values within each row sum to 100%. The analysis focuses on the dominant error patterns across the five misinformation categories and provides additional insight into the class-level results reported in Table 2.

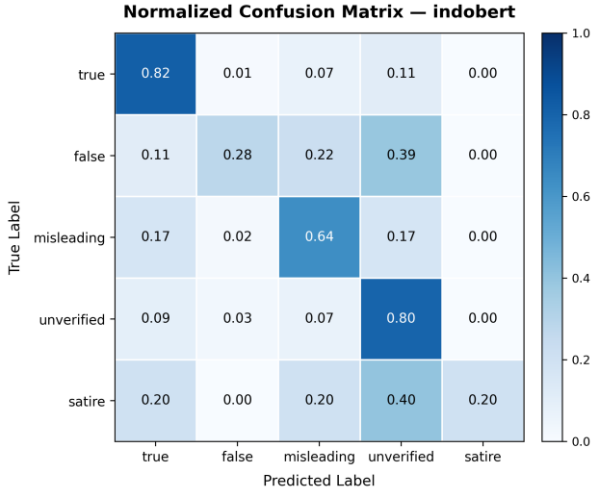

Figure 4. Normalized confusion matrix for IndoBERT.

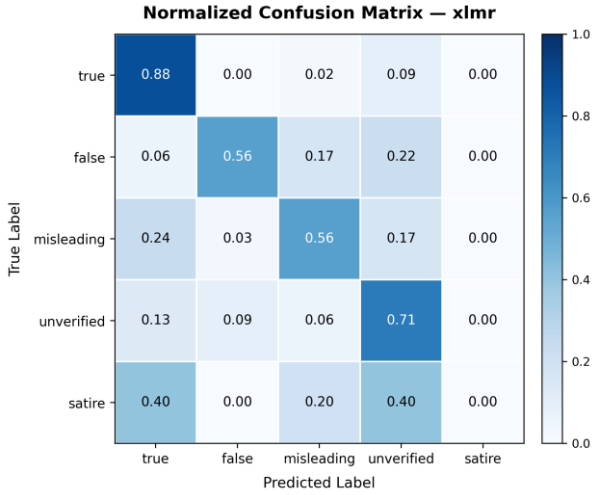

Figure 5. Normalized confusion matrix for XLM-RoBERTa.

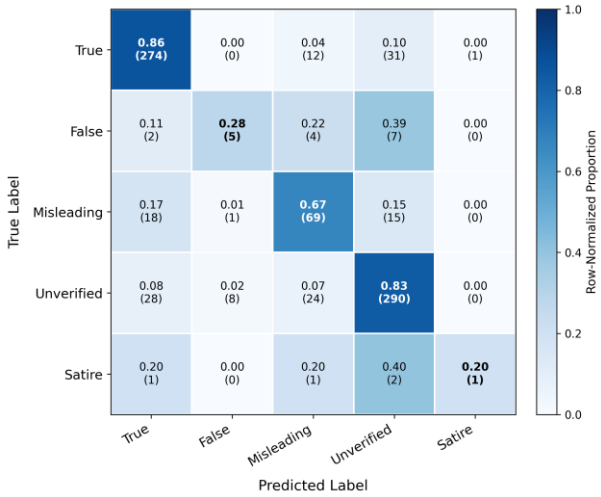

Figure 6. Normalized confusion matrix for the proposed ensemble.

Across the three models, two error patterns are particularly notable. First, a substantial proportion of Misleading posts are classified as True. For the ensemble, 52.9% of the misclassified Misleading samples, corresponding to 18 of 34 errors, are assigned to the True category. This pattern is plausible given the definition of Misleading content, as such posts may contain genuine agricultural facts while introducing an incorrect, exaggerated, or unsupported conclusion. Consequently, semantic overlap with legitimate agricultural information can make these samples difficult to distinguish from genuinely True posts.

Second, False posts are frequently classified as Unverified. For the ensemble, 53.8% of the misclassified False samples, corresponding to 7 of 13 errors, are assigned to the Unverified category. This behavior suggests that the model can recognize uncertainty around a claim but may lack sufficient information to confidently distinguish a demonstrably false claim from one for which evidence is unavailable. Although this tendency may reduce the risk of incorrectly classifying false agricultural claims as True, it also highlights a limitation of text-based classification: distinguishing False from Unverified may require external evidence beyond linguistic information alone.

The error proportions of 52.9% and 53.8% are calculated only among the misclassified samples within the respective true classes and therefore differ from the row-normalized percentages displayed in Figures 4–6, where all samples in each true class are used as the denominator. The confusion matrices were generated from the same test-set predictions used to calculate the per-class F1-scores reported in Table 2.

The comparison across the three confusion matrices further illustrates the differences observed in the per-class results. The ensemble exhibits a more diagonal-dominant pattern for the better-represented categories, particularly True, Misleading, and Unverified, indicating fewer off-diagonal predictions for these classes. XLM-RoBERTa shows substantial off-diagonal mass for Satire, consistent with its zero F1-score for this category in Table 2. IndoBERT provides intermediate behavior, with stronger performance on False and Satire than XLM-RoBERTa but weaker performance than the ensemble on several of the more populated categories.

These confusion patterns also help explain why the overall improvement of the ensemble is relatively modest despite its higher F1-Macro. The ensemble does not eliminate the fundamental ambiguity between Misleading and True or between False and Unverified. Instead, its contribution appears to be concentrated in reducing errors among the better-represented categories while providing complementary information for more difficult cases. This interpretation is consistent with the per-class results in Table 2, where the ensemble achieves its largest gains on True, Misleading, and Unverified rather than on the two smallest classes.

These observations have practical implications for the proposed system. A text classifier can capture linguistic and contextual patterns associated with misinformation, but its ability to determine whether a specific agricultural claim is factually supported remains limited without access to external evidence. The frequent confusion between False and Unverified therefore provides a direct motivation for the Knowledge Verification module introduced in

Section 3.3.2. Rather than treating every uncertain prediction as an isolated classification error, external agricultural evidence can potentially help distinguish claims that are genuinely false from those for which sufficient evidence is unavailable.

These findings indicate that the principal challenge is not simply distinguishing misinformation from factual content, but separating categories that share substantial semantic overlap. In particular, the Misleading category requires the model to recognize when accurate information is combined with an inappropriate conclusion, whereas the False–Unverified boundary may require evidence beyond surface-level linguistic patterns. These observations further support the use of fine-grained classification together with evidence-based verification for agricultural misinformation detection.

4.5. Ablation Study

Table 3 isolates the contribution of Focal Loss and data augmentation using IndoBERT on the validation set (n = 793). All configurations in this table are evaluated on the validation set, while the corresponding test-set results are reported in Table 1.

Table 3. Ablation study on the validation set (IndoBERT base)

Configuration	F1-Macro	Gain vs. Baseline
Cross-entropy, no augmentation (Baseline)	0.4843	—
+ Focal Loss only	0.5491	+0.0648
+ Augmentation only	0.5123	+0.0280
+ Focal Loss and Augmentation	0.5809	+0.0966
Ensemble (Ours)	0.6381	+0.1538

The baseline configuration using standard cross-entropy without augmentation obtains an F1-Macro of 0.4843. Adding Focal Loss increases the score to 0.5491, corresponding to a gain of 6.48 percentage points. Augmentation alone produces a smaller improvement to 0.5123, or 2.80 percentage points. When both techniques are applied, F1-Macro increases to 0.5809, representing a total gain of 9.66 percentage points over the baseline. The combined improvement is slightly larger than the sum of the individual gains (6.48 + 2.80 = 9.28 points), suggesting a complementary interaction between the two techniques rather than independent additive effects [46].

This interaction is consistent with the role of the two components during training. Focal Loss places greater emphasis on difficult examples, but its effectiveness depends on having sufficient minority-class examples from which to learn difficult patterns. The original training set contained only 84 False and 24 Satire samples. Augmentation increases the number and lexical diversity of minority-class examples, thereby providing Focal Loss with a broader set of difficult instances. The results therefore suggest that augmentation and loss reweighting address different aspects of the same imbalance problem.

Adding XLM-RoBERTa as the ensemble partner further increases the validation F1-Macro to 0.6381. This result suggests that the second Transformer provides complementary information beyond what the IndoBERT configuration using Focal Loss and augmentation provides. However, because this value differs from the 0.5870 ± 0.0028 reported for the ensemble in Table 1, the distinction between validation and test results should be maintained explicitly throughout the manuscript.

The ablation results therefore support the inclusion of both Focal Loss and augmentation in the training pipeline, while also indicating that the ensemble contribution should be interpreted separately from the training-component gains. In particular, the validation results demonstrate component-level effects, whereas the test-set results in Table 1 provide the final held-out evaluation.

4.6. Knowledge Verification Analysis

The Knowledge Verification (KV) module is intended to provide evidence-based information that is not directly available from the text classification signal. Its effectiveness is therefore better characterized by its coverage and behavior across different evidence verdicts than by its contribution to a single aggregate classification metric.

Agricultural entities were detected in 74.3% of the test posts (n = 590), indicating that the current entity dictionary and knowledge sources can be applied to a substantial portion of the test set. Among these posts, the retrieved evidence produced three verdicts: supported (38.2%), insufficient_evidence (41.5%), and refuted (20.3%). The relatively large proportion of insufficient evidence indicates that evidence retrieval does not automatically yield a decisive factual verdict for every detected agricultural entity. This is expected for claims that require contextual or domain-specific evidence beyond what is available in current knowledge sources.

The verdict distribution also provides an important indication of how the KV module differs from the five-class classifier. The supported verdict accounts for 38.2% of entity-detected posts, while the refuted verdict accounts for 20.3%. The latter proportion is substantially higher than the False class proportion in the dataset. This difference suggests that the KV module may identify individual incorrect sub-claims within posts that are not necessarily labeled as False at the post level, particularly when a Misleading or Unverified post contains a specific claim that conflicts with a knowledge-base entry. Thus, the KV signal should not be interpreted as a direct replacement for the five-class label.

Two representative cases illustrate this behavior. A claim that urea fertilizer contains 46% nitrogen received a supported verdict with a confidence score of 0.80, consistent with the corresponding entry in the local knowledge base. Conversely, the claim that organic pesticides are always safer than synthetic pesticides received a refuted verdict with a confidence score of 0.70, consistent with the relevant misconception entry. These examples illustrate how retrieved evidence can provide an explicit basis for the system's output, although they do not by themselves establish the general accuracy of the verification module.

The quantitative ablation provides a more direct assessment of the module's contribution. Compared with the ML-only configuration, adding KV increases accuracy by 1.13 percentage points and F1-Weighted by 0.60 points, while F1-Macro decreases by 0.0056 on the full test set. On the subset in which agricultural entities are detected, the reported accuracy improvement reaches 1.8 percentage points. This trade-off indicates that the current KV implementation does not uniformly improve the five-class classification objective. Its contribution is more apparent in aggregate accuracy and evidence generation than in macro-averaged minority-class performance.

One explanation is the mapping between knowledge verdicts and the five-class prediction space. Supported evidence tends to contribute probability mass toward the True category, which is already one of the better-represented classes. In contrast, the current knowledge sources provide less direct information for distinguishing False from Unverified or for recognizing Satire and misleading framing. Consequently, the KV module can provide useful factual evidence without necessarily improving every class in the classification task.

The coverage result also identifies a clear limitation. Approximately one quarter of test posts do not contain entities recognized by the current agricultural dictionary and therefore fall back to the ML-only pathway. In addition, the knowledge base currently covers only a limited set of agricultural entities. Expanding the entity dictionary and developing a broader structured Indonesian agricultural knowledge base could increase coverage, although broader matching would also need to be controlled to avoid false entity detections. These limitations suggest that future improvements should focus not only on the fusion weights but also on retrieval coverage and evidence quality.

4.7. Data Ethics and Governance

The collection and release of social media data require particular attention to privacy and responsible data sharing. Usernames, profile photographs, and direct identifiers were removed before annotation. Data collection was conducted using the YouTube Data API v3 and Apify infrastructure, with the collection configuration respecting the applicable platform requirements. Twitter was disabled in the collection configuration.

The publicly released dataset is limited to cleaned text, assigned labels, and platform category. User identifiers, post URLs, profile links, engagement metrics, and comment-thread structures are excluded from the released dataset. This minimizes unnecessary disclosure while preserving the primary information required for reproducible classification research. Researchers requiring additional metadata may request access through a data-access agreement subject to institutional review requirements.

The annotation process also considered the exposure of human reviewers to potentially misleading or problematic content. GPT-4o-mini was used for weak labeling, while the human validation subset was reviewed by two domain experts who were informed of the research purpose and intended use of the resulting dataset.

These measures provide a reasonable basis for responsible dataset sharing while retaining the information necessary for research replication. Nevertheless, future releases should continue to follow platform-specific data-sharing requirements and ensure that previously public information is not redistributed in a form that unnecessarily increases the risk of identifying individual users.

5. Conclusions

AgriMisinfo-ID provides a bilingual dataset for agricultural misinformation detection in Indonesian social media contexts, comprising 5,288 labeled samples across five categories from YouTube, Instagram, Facebook, and supplementary public sources. The proposed framework combines IndoBERT and XLM-RoBERTa with Focal Loss, paraphrase augmentation, and weighted soft voting, achieving an F1-Macro of 0.5870 ± 0.0028 and an accuracy of 0.8027 ± 0.0039 across three seeds on the test set. The ablation results indicate that Focal Loss and augmentation provide complementary benefits, while the Knowledge Verification module adds evidence-based information to the classification process.

Several limitations remain. False and Satire detection is still unreliable because of the small number of original samples available for these categories. Additional independently validated minority-class data are therefore needed for more reliable evaluation. The Knowledge Verification module also has limited coverage, with 25.7% of test posts lacking entity-based evidence. Expanding the agricultural knowledge base and improving retrieval coverage are important directions for future work. Extending the dataset to additional platforms such as TikTok and WhatsApp would further broaden its coverage. The dataset and associated research resources will be released publicly to support reproducible research in agricultural misinformation detection.

Author Contributions: Conceptualization: B.R.P. D.; Methodology: B.R.P. D.; Software: B.R.P. D. and D.B. F.; Validation: N.O. A. and D.B. F.; Formal analysis: B.R.P. D.; Investigation: N.O. A. and F.A. H.; Data curation: N.O. A., F.A. H., and N.P.H. J.; Writing—original draft preparation: B.R.P. D.; Writing—review and editing: B.R.P. D.; Visualization: N.P.H. J.; Project administration: F.Y. P.; Supporting data: T. P. and F.Y. P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Dana Hibah Penelitian dan Pengabdian Kepada Masyarakat Universitas Jember Tahun Anggaran 2026 under contract number 4100/DST/UN25.C1/LT/2026, scheme Hibah Keris-Dimas (Kelompok Riset dan Pengabdian Masyarakat).

Data Availability Statement: Dataset available at https://huggingface.co/datasets/brianrizqi/agri-misinformation.

Acknowledgments: This work was carried out under the Hibah Keris-Dimas (Kelompok Riset dan Pengabdian Masyarakat) scheme of Universitas Jember, supported by Dana Hibah Penelitian dan Pengabdian Kepada Masyarakat Universitas Jember Tahun Anggaran 2026 with contract number 4100/DST/UN25.C1/LT/2026.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake News Detection on Social Media," *ACM SIGKDD Explor. Newsl.*, vol. 19, no. 1, pp. 22–36, Sep. 2017, doi: 10.1145/3137597.3137600.

[2] E. Aïmeur, S. Amri, and G. Brassard, "Fake news, disinformation and misinformation in social media: a review," *Soc. Netw. Anal. Min.*, vol. 13, no. 1, p. 30, Feb. 2023, doi: 10.1007/s13278-023-01028-5.

[3] M. R. Islam, S. Liu, X. Wang, and G. Xu, "Deep learning for misinformation detection on online social networks: a survey and new perspectives," *Soc. Netw. Anal. Min.*, vol. 10, no. 1, p. 82, Dec. 2020, doi: 10.1007/s13278-020-00696-x.

[4] A. Rahmawati, A. Alamsyah, and A. Romadhony, "Hoax News Detection Analysis using IndoBERT Deep Learning Methodology," in *2022 10th International Conference on Information and Communication Technology (ICoICT)*, Aug. 2022, pp. 368–373. doi: 10.1109/ICoICT55009.2022.9914902.
[5] B. P. Nayoga, R. Adipradana, R. Suryadi, and D. Suhartono, "Hoax Analyzer for Indonesian News Using Deep Learning Models," *Procedia Comput. Sci.*, vol. 179, pp. 704–712, 2021, doi: 10.1016/j.procs.2021.01.059.
[6] A. Chowdhury, K. H. Kabir, A.-R. Abdulai, and M. F. Alam, "Systematic Review of Misinformation in Social and Online Media for the Development of an Analytical Framework for Agri-Food Sector," *Sustainability*, vol. 15, no. 6, p. 4753, Mar. 2023, doi: 10.3390/su15064753.
[7] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in *Proceedings of the 28th International Conference on Computational Linguistics*, Nov. 2020, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.
[8] A. Conneau *et al.*, "Unsupervised Cross-lingual Representation Learning at Scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8440–8451. doi: 10.18653/v1/2020.acl-main.747.
[9] V. U. Gongane, M. V. Munot, and A. D. Anuse, "A survey of explainable AI techniques for detection of fake news and hate speech on social media platforms," *J. Comput. Soc. Sci.*, vol. 7, no. 1, pp. 587–623, Apr. 2024, doi: 10.1007/s42001-024-00248-9.
[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020, doi: 10.1109/TPAMI.2018.2858826.
[11] J. C. S. Reis, A. Correia, F. Murai, A. Veloso, and F. Benevenuto, "Supervised Learning for Fake News Detection," *IEEE Intell. Syst.*, vol. 34, no. 2, pp. 76–81, Mar. 2019, doi: 10.1109/MIS.2019.2899143.
[12] J. Devlin, M.-W. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North*, Oct. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.
[13] R. K. Kaliyar, A. Goswami, and P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," *Multimed. Tools Appl.*, vol. 80, no. 8, pp. 11765–11788, Mar. 2021, doi: 10.1007/s11042-020-10183-2.
[14] 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. 16, p. e35865, Aug. 2024, doi: 10.1016/j.heliyon.2024.e35865.
[15] S. A. Aljawarneh and S. A. Swedat, "Fake News Detection Using Enhanced BERT," *IEEE Trans. Comput. Soc. Syst.*, vol. 11, no. 4, pp. 4843–4850, Aug. 2024, doi: 10.1109/TCSS.2022.3223786.
[16] E. Essa, K. Omar, and A. Alqahtani, "Fake news detection based on a hybrid BERT and LightGBM models," *Complex Intell. Syst.*, vol. 9, no. 6, pp. 6581–6592, Dec. 2023, doi: 10.1007/s40747-023-01098-0.
[17] V.-I. Ilie, C.-O. Truica, E.-S. Apostol, and A. Paschke, "Context-Aware Misinformation Detection: A Benchmark of Deep Learning Architectures Using Word Embeddings," *IEEE Access*, vol. 9, pp. 162122–162146, 2021, doi: 10.1109/ACCESS.2021.3132502.
[18] S.-Y. Lin, Y.-C. Kung, and F.-Y. Leu, "Predictive intelligence in harmful news identification by BERT-based ensemble learning model with text sentiment analysis," *Inf. Process. Manag.*, vol. 59, no. 2, p. 102872, Mar. 2022, doi: 10.1016/j.ipm.2022.102872.
[19] M. Al-alshaqi, D. B. Rawat, and C. Liu, "Ensemble Techniques for Robust Fake News Detection: Integrating Transformers, Natural Language Processing, and Machine Learning," *Sensors*, vol. 24, no. 18, p. 6062, Sep. 2024, doi: 10.3390/s24186062.
[20] M. E.Almandouh, M. F. Alrahmawy, M. Eisa, M. Elhoseny, and A. S. Tolba, "Ensemble based high performance deep learning models for fake news detection," *Sci. Rep.*, vol. 14, no. 1, p. 26591, Nov. 2024, doi: 10.1038/s41598-024-76286-0.
[21] M. S. Khan and C. Hongsong, "Hybrid transformer deep neural architectures for enhanced misinformation detection on social media," *Expert Syst. Appl.*, vol. 300, p. 130470, Mar. 2026, doi: 10.1016/j.eswa.2025.130470.
[22] C. Kumar, M. Bansal, M. A. Khan, V. Kaushik, M. Arquam, and A. Alabdultif, "Graph-augmented transformer ensemble framework for robust and scalable fake news detection in social media ecosystems," *Sci. Rep.*, vol. 16, no. 1, p. 2001, Dec. 2025, doi: 10.1038/s41598-025-31653-3.
[23] M. F. Mridha, A. J. Keya, M. A. Hamid, M. M. Monowar, and M. S. Rahman, "A Comprehensive Review on Fake News Detection With Deep Learning," *IEEE Access*, vol. 9, pp. 156151–156170, 2021, doi: 10.1109/ACCESS.2021.3129329.
[24] D. Plikynas, I. Rizgelienė, and G. Korvel, "Systematic Review of Fake News, Propaganda, and Disinformation: Examining Authors, Content, and Social Impact Through Machine Learning," *IEEE Access*, vol. 13, pp. 17583–17629, 2025, doi: 10.1109/ACCESS.2025.3530688.
[25] N. Seddari, A. Derhab, M. Belaoued, W. Halboob, J. Al-Muhtadi, and A. Bouras, "A Hybrid Linguistic and Knowledge-Based Analysis Approach for Fake News Detection on Social Media," *IEEE Access*, vol. 10, pp. 62097–62109, 2022, doi: 10.1109/ACCESS.2022.3181184.
[26] B. Xie, X. Ma, J. Wu, J. Yang, and H. Fan, "Knowledge Graph Enhanced Heterogeneous Graph Neural Network for Fake News Detection," *IEEE Trans. Consum. Electron.*, vol. 70, no. 1, pp. 2826–2837, Feb. 2024, doi: 10.1109/TCE.2023.3324661.
[27] G. Joshi *et al.*, "Explainable Misinformation Detection Across Multiple Social Media Platforms," *IEEE Access*, vol. 11, pp. 23634–23646, 2023, doi: 10.1109/ACCESS.2023.3251892.
[28] S. Kuntur, A. Wróblewska, M. Paprzycki, and M. Ganzha, "Under the Influence: A Survey of Large Language Models in Fake News Detection," *IEEE Trans. Artif. Intell.*, vol. 6, no. 2, pp. 458–476, Feb. 2025, doi: 10.1109/TAI.2024.3471735.
[29] Muhammad Ikram Kaer Sinapoy, Yuliant Sibaroni, and Sri Suryani Prasetyowati, "Comparison of LSTM and IndoBERT Method in Identifying Hoax on Twitter," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 7, no. 3, pp. 657–662, Jun. 2023, doi: 10.29207/resti.v7i3.4830.
[30] T. C. Praha, W. Widodo, and M. Nugraheni, "Indonesian Fake News Classification Using Transfer Learning in CNN and LSTM," *JOIV Int. J. Informatics Vis.*, vol. 8, no. 3, p. 1213, Sep. 2024, doi: 10.62527/joiv.8.2.2126.
[31] A. D. Safira and E. B. Setiawan, "Hoax Detection in Social Media using Bidirectional Long Short-Term Memory (Bi-LSTM) and 1 Dimensional-Convolutional Neural Network (1D-CNN) Methods," in *2023 11th International Conference on Information and Communication Technology (ICoICT)*, Aug. 2023, pp. 355–360. doi: 10.1109/ICoICT58202.2023.10262528.

[32] L. A. Pekandi, R. G. Widjaja, A. Ananta, J. Harefa, and K. Jingga, "Evaluating IndoBERT for Indonesian Hoax News Detection: A Comparative Study with Ensemble and CNN-LSTM Models," *Procedia Comput. Sci.*, vol. 269, pp. 1625–1633, 2025, doi: 10.1016/j.procs.2025.09.105.

[33] Y. Sagama and A. Alamsyah, "Multi-Label Classification of Indonesian Online Toxicity using BERT and RoBERTa," in *2023 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, Jul. 2023, pp. 143–149. doi: 10.1109/IAICT59002.2023.10205892.

[34] J. Sirusstara, N. Alexander, A. Alfarisy, S. Achmad, and R. Sutoyo, "Clickbait Headline Detection in Indonesian News Sites using Robustly Optimized BERT Pre-training Approach (RoBERTa)," in *2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, Sep. 2022, pp. 1–6. doi: 10.1109/AiDAS56890.2022.9918678.

[35] J. Laimeheriwa, I. A. Iswanto, W. Oktriani, and M. F. Hidayat, "A Survey on Indonesian Hoax Analyzer and Fake News Detection Using Deep Learning Techniques," in *2024 6th International Conference on Cybernetics and Intelligent System (ICORIS)*, Nov. 2024, pp. 1–6. doi: 10.1109/ICORIS63540.2024.10903927.

[36] D. B. Firmawan and B. R. P. Darnoto, "Cross-Domain Faithfulness Evaluation of SHAP and Attention-Based Explanations in Transformer NLP Models," *J. Comput. Theor. Appl.*, vol. 4, no. 1, pp. 146–163, Jul. 2026, doi: 10.62411/jcta.16258.

[37] B. R. P. Darnoto and D. B. Firmawan, "Language-Similarity-Guided Transfer Fine-Tuning of Pre-trained Transformer Models for Sentiment Analysis Across 12 Indonesian Regional Languages," *J. Comput. Theor. Appl.*, vol. 3, no. 4, pp. 547–563, May 2026, doi: 10.62411/jcta.15975.

[38] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: a Large-scale Dataset for Fact Extraction and VERification," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 809–819. doi: 10.18653/v1/N18-1074.

[39] J. Kim, S. Park, Y. Kwon, Y. Jo, J. Thorne, and E. Choi, "FactKG: Fact Verification via Reasoning on Knowledge Graphs," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 16190–16206. doi: 10.18653/v1/2023.acl-long.895.

[40] Z. Chen, Y. Wang, B. Zhao, J. Cheng, X. Zhao, and Z. Duan, "Knowledge Graph Completion: A Review," *IEEE Access*, vol. 8, pp. 192435–192456, 2020, doi: 10.1109/ACCESS.2020.3030076.

[41] A. Golda *et al.*, "Privacy and Security Concerns in Generative AI: A Comprehensive Survey," *IEEE Access*, vol. 12, pp. 48126–48144, 2024, doi: 10.1109/ACCESS.2024.3381611.

[42] G. G. Şahin, "To Augment or Not to Augment? A Comparative Study on Text Augmentation Techniques for Low-Resource NLP," *Comput. Linguist.*, vol. 48, no. 1, pp. 5–42, Apr. 2022, doi: 10.1162/coli_a_00425.

[43] F. Muftie and M. Haris, "IndoBERT Based Data Augmentation for Indonesian Text Classification," in *2023 International Conference on Information Technology Research and Innovation (ICITRI)*, Aug. 2023, pp. 128–132. doi: 10.1109/ICITRI59340.2023.10250061.

[44] F. Gilardi, M. Alizadeh, and M. Kubli, "ChatGPT outperforms crowd workers for text-annotation tasks," *Proc. Natl. Acad. Sci.*, vol. 120, no. 30, Jul. 2023, doi: 10.1073/pnas.2305016120.

[45] X. He *et al.*, "AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, 2024, pp. 165–190. doi: 10.18653/v1/2024.naacl-industry.15.

[46] L. Rokach, "Ensemble-based classifiers," *Artif. Intell. Rev.*, vol. 33, no. 1–2, pp. 1–39, Feb. 2010, doi: 10.1007/s10462-009-9124-7.

[47] D. Jiang and J. He, "Text Semantic Classification of Long Discourses Based on Neural Networks with Improved Focal Loss," *Comput. Intell. Neurosci.*, vol. 2021, no. 1, Jan. 2021, doi: 10.1155/2021/8845362.

[48] H. R. Saeidnia, E. Hosseini, B. Lund, M. A. Tehrani, S. Zaker, and S. Molaei, "Artificial intelligence in the battle against disinformation and misinformation: a systematic review of challenges and approaches," *Knowl. Inf. Syst.*, vol. 67, no. 4, pp. 3139–3158, Apr. 2025, doi: 10.1007/s10115-024-02337-7.