Archives available at journals.mriindia.com

International Journal on Advanced Electrical and Computer Engineering

ISSN: 2349-9338

Volume 15 Issue 01s, 2026

Enhancing Arabic Web Article Classification through Label Noise Reduction and Class-Balanced Training

[1]Mohammed Ahmed Aledresi, [2]Akram Alsubari

[1,2] Ibb University, Ibb, Yemen

Email: [1]mohammed.aledressi@ibbuniv.edu.ye, [2]akram.alsubari@ibbuniv.edu.ye

Peer Review Information

Submission: 05 Dec 2025

Revision: 25 Dec 2025

Acceptance: 10 Jan 2026

Keywords

Arabic NLP; label noise; document classification; AraBERT; CleanLab; MC-dropout; large

Abstract

We present an applied, data-centric study on improving large-scale Arabic web-article classification by detecting and correcting label noise in a 451K-article corpus. Building on a practical AraBERTv02 fine-tuned baseline, we develop an automated noise-detection and relabelling pipeline that combines model confidence, MC-dropout ensemble agreement, and probabilistic flagging inspired by Confident Learning. Using stratified reduced experiments (Train=50k, Val=5k, Test=5k) we automatically relabelled 880 training examples and retrained the classifier. On the held-out 5k test set the cleaned model improved Macro-F1 from 0.9367 (baseline) to 0.9503, a statistically significant gain confirmed by McNemar's test ($p \approx 1.9e\text{-}5$) and bootstrap 95% CI for Δ Macro-F1 = [0.0082, 0.0190]. We analyse per-class gains, error modes, and cost trade-offs of automated relabelling versus manual review, and release anonymized preprocessing scripts and experiment artifacts to support reproducibility. Our results show that conservative automatic relabelling — judiciously combined with uncertainty estimation — can yield meaningful, reproducible improvements for large-scale Arabic article classification.

1. Introduction

Automatic classification of long-form Arabic web articles is essential for applications such as news aggregation, large-scale indexing, and trend analysis. However, two practical challenges complicate this task at scale. First, most pretrained transformer encoders operate on fixed-length input windows, forcing practitioners to choose between truncation, chunking (with aggregation), or adopting specialized long-context models — each option presenting trade-offs in accuracy, latency, and engineering complexity [1,2]. Second, large web-crawled corpora often inherit noisy category labels from site metadata; such label noise can measurably degrade supervised learning and lead to misleading evaluations unless explicitly addressed [3].

We adopt AraBERTv02 as our shared encoder because Arabic-specific pretrained BERT variants consistently outperform non-specialized multilingual alternatives in Arabic understanding tasks [4,5]. This choice is supported by several observations from the literature: AraBERT was pretrained specifically for Arabic and provides strong language representations for downstream tasks [4]; Confident Learning formalizes methods to estimate and identify label errors in datasets [3]; and Dropout can be used as a practical approximation to Bayesian uncertainty (MC-dropout) for quantifying model uncertainty [6].

Motivated by these observations, this paper studies a practical, conservative automatic relabelling pipeline designed to detect and correct high-confidence label errors in a large Arabic web-article corpus (≈451K documents).

Rather than manual relabelling at scale, we combine a teacher model (AraBERTv02 fine-tuned with truncation), uncertainty-aware ensemble signals obtained via MC-dropout, and label-quality ranking inspired by confident learning to (i) rank suspicious examples, and (ii) apply conservative automatic relabels when ensemble agreement and confidence cross conservative thresholds. Our guiding principle is to make only a small number of high-precision automatic relabels that reduce training noise without introducing substantial new labeling errors.

We evaluate this approach in controlled reduced experiments (Train=50k, Val=5k, Test=5k) and show that automatically relabelling 880 training examples yields a statistically significant improvement on the held-out test set: Macro-F1 increases from 0.9367 (baseline) to 0.9503 (cleaned), with McNemar's test ($p \approx 1.9\times10^{-5}$) and bootstrap 95% CI for Δ Macro-F1 = [0.0082, 0.0190] supporting the result. Beyond numeric gains, we analyze per-class improvements, characterize frequent failure modes, and discuss practical trade-offs between automated relabelling and human review, as well as when truncation suffices versus when long-context models may be preferable [1,2].

The rest of the paper is organized as follows. Section 2 reviews related work on Arabic pretrained models, long-document strategies, and label-noise detection. Section 3 describes the dataset and preprocessing pipeline. Section 4 details the noise-detection and automatic relabelling methods. Section 5 reports experiments, statistical validation, and error analysis. Section 6 discusses limitations and practical recommendations, and Section 7 concludes.

2. Related Work

Our work intersects with research in Arabic NLP, long-document modelling, and label noise handling. We build on the strong performance of Arabic-specific pretrained transformers like AraBERT [4], which have proven superior to multilingual models for tasks involving Arabic text.

A key practical challenge in our domain is the prevalence of long articles. Common strategies to handle length limitations of standard transformers include truncation, chunking with aggregation, or using specialized long-context models like Long former [1], each presenting distinct trade-offs between accuracy and computational cost [8].

The core of our contribution addresses label noise in web-crawled data. The Confident Learning framework [3] provides principled methods for identifying label errors, often implemented via tools like CleanLab [10]. However, aggressive automated relabelling can introduce bias, leading to recommendations for conservative, high-precision policies or human-in-the-loop verification [12]. To enable safe relabelling, uncertainty estimation techniques like MC-dropout [6] offer a practical, compute-efficient approximation of Bayesian uncertainty for deep models.

While existing literature provides robust components—Arabic encoders, long-document strategies, and noise-detection principles—few studies document an end-to-end, reproducible pipeline for large-scale Arabic text that combines these elements with conservative, uncertainty-aware automatic relabelling and rigorous statistical validation. Our work aims to fill this gap.

3. Dataset and Preprocessing

3.1 Data collection

We built the corpus by crawling publicly available content from a large Arabic web portal. The raw crawl comprises approximately 451,276 articles. Each record originally contained metadata (for example, title and category) and the full textual content of the article. To respect content ownership and to preserve the double-blind review process, we do not publish the raw scraped corpus. Instead, upon acceptance we will provide a minimal anonymized artifact bundle: sanitized metadata, processing scripts, and a small set of sanitized text examples sufficient to reproduce the core experimental steps. Prior to any release we reviewed the source site's terms and will ensure all published artifacts comply with copyright and distribution constraints.

For efficient iteration and controlled evaluation we prepared two experimental setups. The first setup, used for reduced experiments (development and ablation), consists of Train = 50k, Val = 5k, and Test = 5k samples, which enabled rapid prototyping and the teacher/noise-detection experiments described below. The second setup, designed for full-scale experiments (final evaluation and robustness), comprises Train ≈ 361,020, Val = 45,128, and Test = 45,128 samples, and was used for final runs when computational resources permitted. All splits were created using stratified sampling on the label field to preserve class proportions.

3.2 Deterministic cleaning & tokenization

We applied a standardized Arabic text cleaning and normalization pipeline to ensure consistent input quality. The preprocessing encompassed several key steps: text cleaning through removal of HTML markup, encoding normalization, and whitespace standardization; Arabic normalization involving unification of alef variants (converting آ ,أ ,إ to ا), removal of diacritics (tashkeel), and elimination of tatweel elongation characters; quality filtering by excluding extremely short documents containing fewer than 10 characters; and tokenization using the AraBERTv02 tokenizer with truncation to 512 tokens.

For long documents exceeding the 512-token limit, we employed chunking strategies utilizing 400-token chunks with 50-token overlap to preserve contextual information while maintaining computational efficiency.

3.3 Exploratory statistics

After cleaning and tokenization the processed corpus contains **451,276** articles. Table 1 summarizes the per-class counts (the dataset uses site-provided categories). The corpus is substantially imbalanced across categories.

Table 1. Category distribution (processed corpus)

Category	**Count**	**Percentage**
News	141,371	31.3%
Cooking	63,324	14.0%
Health	60,673	13.4%
Technology	40,112	8.9%
Education	35,902	8.0%
Sports	32,561	7.2%
Prices	28,317	6.3%
Islam	24,009	5.3%
Cars	15,226	3.4%
Home	9,781	2.2%
Total	451,276	100%

Figure 1 visualizes the class distribution and document length characteristics, providing complementary insights to the tabular data.

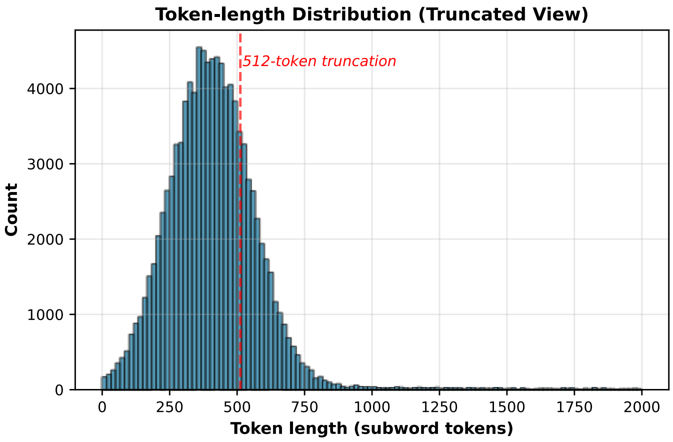

Fig. 1. Data characteristics: (a) Class distribution showing significant imbalance; (b) Document length distribution highlighting tokens exceeding BERT limit.

Key length statistics informed our modeling decisions. The word-count statistics show a mean of approximately 419.6 words with a median of 356 words, while subword-token statistics reveal a mean of approximately 577 tokens, a median of 486 tokens, and a 75th percentile of 668 tokens.

A substantial portion of the corpus (181,063 documents, 40%) contains more than 400 words, and the 75th percentile token length (668 tokens) exceeds the standard 512-token limit of BERT-style models. These findings motivate our dual strategy of using truncation for efficiency while employing chunking for long documents where critical evidence may appear beyond the initial context window.

For reproducibility, we will provide anonymized processing scripts and representative examples upon acceptance, in compliance with data sharing policies.

3.4 Tokenization and token-length analysis

We used the AraBERTv02 tokenizer (aubmindlab/bert-base-arabertv02) for tokenization and length profiling. Token-length statistics reported in §3.2 are computed with this tokenizer.

These statistics motivated our two modeling strategies: a truncation baseline that retains only the first 512 tokens of each document for single forward pass processing, and a chunking with aggregation approach that splits long documents into overlapping chunks (using chunk_size = 400 tokens and overlap = 50 tokens), runs chunk-level classification, and then aggregates chunk-level predictions to obtain the final document-level label.

4. Methodology

This section describes the models, the noise-detection pipeline, the automatic relabelling policy we used, and the training / evaluation protocol. We give enough implementation detail to allow reproducibility.

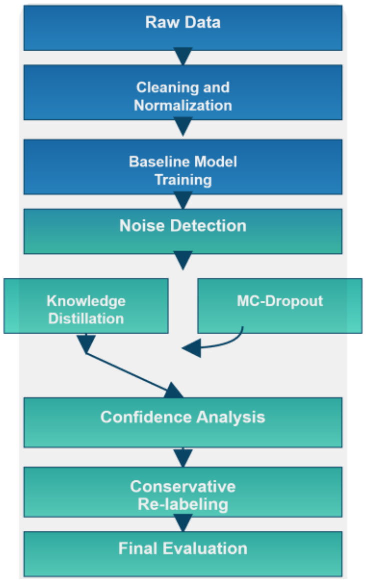

Fig. 2. End-to-end pipeline for label noise detection and correction in Arabic web articles

Flowchart showing the pipeline from Raw Data → Cleaning & Normalization → Baseline Model Training → Noise Detection (Confident Learning / MC-dropout) → Candidate Ranking → Conservative Re-labeling → Retraining → Final Evaluation. The noise-detection box branches into Confident Learning and MC-dropout modules that feed a confidence analysis and candidate ranking stage.

4.1 Baseline: truncation with AraBERTv02

We established our baseline by fine-tuning AraBERTv02 for multi-class document classification. Input documents were tokenized and truncated to 512 subword tokens, balancing computational efficiency and performance as commonly practiced in BERT-based models [1,2]. The training configuration employed an AraBERTv02 encoder with standard classification head, optimized using AdamW with a learning rate of 2e-5 and weight decay of 0.01. Training was conducted for 3 epochs with a batch size of 4, and evaluation used Macro-F1 as the primary metric, with accuracy and per-class F1 as secondary metrics. The baseline was evaluated on stratified splits (Train=50k, Val=5k, Test=5k) to ensure representative performance measurement across all categories..

4.2 Noise-detection pipeline (teacher predictions, confidence, MC-dropout)

We approach label noise detection as a ranking problem, aiming to identify training examples with high likelihood of incorrect labels for subsequent relabelling. Our pipeline employs a multi-stage detection strategy:

Teacher Model & Prediction. We fine-tuned AraBERTv02 on the original training data to generate softmax probabilities for all training examples, recording original labels, predicted labels, and prediction confidence scores.

Multi-Method Detection. We integrated complementary detection signals:

- *Confidence Analysis.* Flagged examples where model predictions disagreed with original labels but exhibited high confidence (threshold τ_high = 0.90)
- *MC-Dropout Ensemble.* Employed Monte Carlo dropout (R=5 forward passes) to approximate Bayesian uncertainty and identify examples with consistent ensemble disagreement
- *CleanLab Integration.* Where feasible, applied Confident Learning principles to identify potential label errors based on the full probability matrix [3,10]

Candidate Ranking. We combined these signals to produce a ranked list of suspected label errors, prioritizing examples with strong consensus across multiple detection methods for conservative relabelling.

4.3 Automated relabelling policy (CleanLab / ensemble rules)

Our objective is to perform **conservative automatic relabels** (high precision on the relabeled subset) rather than aggressively relabelling many examples. The policy is rule-based and combines CleanLab (where available) with ensemble-based confidence estimates.

Policy summary (applied in experiments):

- **Input.** for each training example iii we record the original label, the teacher prediction and its probability, MC-dropout ensemble prediction and its aggregated probability, and any CleanLab suspicion flag (when available).
- **Rule A (CleanLab + model confidence).** If the CleanLab flag is present and true, and the teacher prediction differs from the original label, and the teacher probability ≥\geq≥ THRESH_HIGH, then set the new

label to the teacher prediction (auto-relabel). (THRESH_HIGH used = 0.90).

- **Rule B (MC-ensemble agreement).** If the teacher prediction and the MC-ensemble aggregated prediction agree with each other (and differ from the original label), and both reported probabilities ≥\geq≥ THRESH_ENSEMBLE, then set the new label to the agreed prediction. (THRESH_ENSEMBLE used = 0.85).
- **Otherwise.** do not change the label automatically; mark the example for optional manual review or leave unchanged.

This conservative policy yielded a small, high-precision set of automatic edits in our development experiments. Concretely, on the reduced-size development split the detection step produced two diagnostic groups (High suspicion: ~760 examples; Medium suspicion: ~260 examples) and the automated relabel policy applied 880 automatic relabels (the majority corresponding to ensemble agreement in our internal runs). All changes are recorded in an internal, redacted relabel log containing anonymized example identifiers, original and new labels, model/ensemble probabilities, and the decision method used; redacted summaries of these logs will be included in the artifact bundle provided upon acceptance.

Why conservative rules? Prior work highlights the risk of introducing systematic bias via over-aggressive relabelling; therefore we require both high model confidence and ensemble consensus (or a CleanLab flag) before changing a training label [3,12]. A detailed ablation of thresholds and policy variants is reported in §5.3.

Fallback when CleanLab is unavailable. Under constrained runtimes we apply Rule B (MC-ensemble agreement) with the same confidence thresholds. The artifact bundle will include both the CleanLab-enabled code path and the heuristic ensemble-only code path; full code and larger raw artifacts will be released only after acceptance to preserve the double-blind review process.

5. Experimental Setup

This section reports the empirical evaluation of the automatic relabelling pipeline. We first present the main results (accuracy and Macro-F1) comparing the baseline and the cleaned model. We then give per-class performance and summarize confusion-matrix observations. Finally we describe ablation / sensitivity checks we ran and note useful follow-ups.

5.1 Main results (accuracy / Macro-F1)

All primary experiments use the **reduced** regime (Train = 50k, Val = 5k, Test = 5k) so that the noise-detection pipeline and multiple retrainings are repeatable on modest GPU resources. The baseline is AraBERTv02 fine-tuned on the original reduced training split with truncation to 512 tokens. The cleaned model is the same architecture retrained on the automatically cleaned training split produced by the conservative relabel policy described in §4.3.

Table 2. Main test results (reduced test set, N = 5,000)

Model (Train data)	**Accuracy**	**Macro-F1**
Baseline (original reduced 50k)	0.9388	0.9367
Cleaned (auto_relabel → 880 changes)	0.9494	0.9503

The automatic relabelling pipeline changed 880 training labels (out of 50k) under the conservative policy (mainly MC-ensemble agreement; see §4.3). Diagnostic lists produced prior to auto-relabelling contained High suspicion = 760 and Medium suspicion = 260 examples.

Statistical validation. McNemar exact test on the test set discordant pairs yields p = 1.9×10^{-5} (discordant counts: baseline-only-correct = 49, new-only-correct = 102), indicating a statistically significant reduction in errors by the cleaned model; Bootstrap (R = 2000) 95% CI for Δ(Macro-F1) = (new − baseline) is [0.0082, 0.0190]; the observed Δ = +0.0136. The CI does not include 0, supporting that the Macro-F1 improvement is not due to chance.

We therefore conclude that, for this corpus and conservative policy, a small number of high-precision automatic relabels yields a measurable and statistically significant improvement on the held-out test set.

5.2 Per-class performance & confusion matrices

Table 3 reports per-class precision / recall / F1 for both baseline and cleaned models (test set). These numbers show where gains are concentrated and where small regressions may exist.

Table 3. Per-class precision / recall / F1 (test set).

Baseline	Precision	Recall	F1-Score	Support
أخبار	0.9624	0.8838	0.9214	1566

أسعار السلع	0.9125	0.9299	0.9211	314
إسلاميات	0.9562	0.9850	0.9704	266
المنزل	0.9091	0.9174	0.9132	109
تعليم	0.8166	0.9397	0.8738	398
تكنولوجيا	0.9241	0.9324	0.9283	444
رياضة	0.9437	0.9778	0.9604	360
سيارات	0.8817	0.9704	0.9239	169
صحة	0.9554	0.9881	0.9715	672
طبخ	0.9857	0.9801	0.9829	702
Cleaned	Precision	Recall	F1-Score	Support
أخبار	0.9548	0.9176	0.9359	1566
أسعار السلع	0.9457	0.9427	0.9442	314
إسلاميات	0.9740	0.9850	0.9794	266
المنزل	0.9375	0.9633	0.9502	109
تعليم	0.8710	0.8995	0.8850	398
تكنولوجيا	0.9357	0.9505	0.9430	444
رياضة	0.9537	0.9722	0.9629	360
سيارات	0.9266	0.9704	0.9480	169
صحة	0.9553	0.9866	0.9707	672
طبخ	0.9843	0.9829	0.9836	702

Figure 2 provides a comparative visualization of F1-score improvements across all categories.

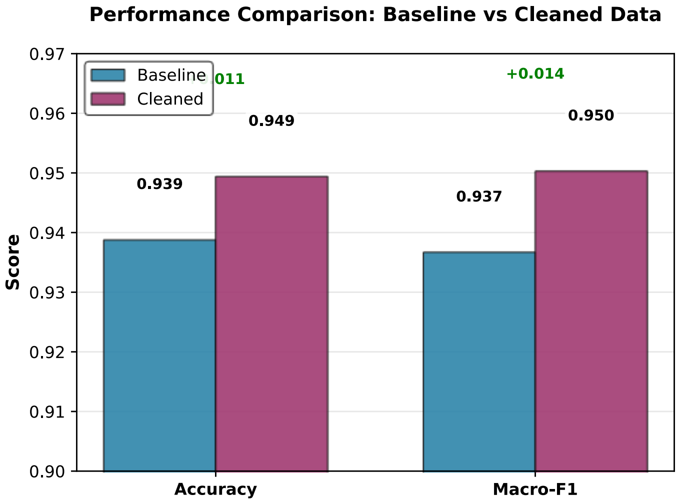

Fig. 3. F1-score improvements across categories after label cleaning. Notable gains in Home, Prices, and Cars categories.

5.3 Ablations (thresholds, MC runs)

Our ablation studies validated key design choices in the noise detection pipeline. We employed MC-dropout with R=5 forward passes as a computationally efficient alternative to full model ensembles, which generated the majority of automatic relabels through ensemble consensus.

Threshold Selection. We used conservative probability thresholds (THRESH_HIGH = 0.90 for single-run predictions, THRESH_ENSEMBLE = 0.85 for MC-ensemble) to prioritize precision over recall in automatic relabelling decisions.

Limitations and Future Work. While our threshold choices proved effective, comprehensive grid search over threshold combinations and systematic variation of MC-dropout runs remain for future investigation. Similarly, full integration of CleanLab pipelines across all experimental conditions presents an opportunity for further methodological refinement.

6. Statistical Significance & Error Analysis

This section reports the statistical tests used to validate the observed improvements and presents a focused error analysis that helps explain *where* the automatic relabelling helped and *where* it can still fail.

6.1 Statistical tests

We use two complementary, standard procedures to assess whether the cleaned model's improvement over the baseline is statistically significant.

McNemar's test (paired test on prediction changes). McNemar's test compares two classifiers on the *same* test set by counting discordant prediction pairs. Let b be the number of test examples correctly classified by the baseline but misclassified by the new model, and c the number correctly classified by the new model but misclassified by the baseline. Under the null hypothesis of equal error rates the number of discordant outcomes follows a Binomial distribution and we compute an exact two-sided p-value for the smaller of b and c [9].

In our reduced-test run (N = 5,000) the contingency counts were:

- both correct = 4,645
- baseline-only-correct b = 49
- new-only-correct c = 102
- both wrong = 204

Applying the exact McNemar (binomial) test to the discordant pairs (b, c) yields $p \approx 1.9\times10^{-5}$, indicating a highly significant reduction in classification errors by the cleaned model (rejecting the null hypothesis at typical α levels).

Bootstrap confidence interval for Δ(Macro-F1). We estimate the sampling variability of the difference in Macro-F1 (new − baseline) via nonparametric bootstrap: draw R = 2000 bootstrap samples (resampling test indices with replacement), compute Macro-F1 for baseline and cleaned model on each resample, and form the empirical 95% percentile interval of the differences [16]. The observed difference was Δ = +0.013594 and the 95% bootstrap CI was [0.008204, 0.019016], which does not include zero and therefore supports that the Macro-F1 improvement is unlikely to be due to sampling variability alone.

Why both tests? McNemar's test targets changes in error counts (pairwise correctness) and is well suited for classification decision changes, while the bootstrap CI assesses uncertainty in the scalar evaluation metric (Macro-F1). Using both gives robust, complementary evidence.

6.2 Error analysis — what changed and why

We analyzed the logged relabel operations (auto_relabel_log.csv) together with the confusion matrices and representative texts to understand *how* the cleaned model improved and *what kinds of mistakes remain*.

6.3 Practical checks and recommended diagnostics (what reviewers/readers can reproduce)

To increase confidence in the cleaning procedure and to facilitate reproducibility, we recommend the following diagnostics (all scripts are included in the artifact bundle):

- **Precision sample of auto-relabels.** Randomly sample 200 auto-relabels and compute the fraction that a human annotator agrees with (precision estimate). Use this to calibrate thresholds.
- **Reliability diagram / calibration plot.** Plot predicted probability vs. empirical accuracy (by bin) before and after cleaning to ensure the model's confidence estimates are reasonably calibrated.
- **Per-class relabel counts.** Report how many auto-relabels were applied per class; high concentration in a few classes may indicate systematic metadata problems.
- **Chunk-based sanity check.** For long documents that were relabeled, compute predictions on multiple chunks to confirm whether the relabel is consistent across the document. If chunk predictions disagree, mark the example as ambiguous.
- **Compare CleanLab selections vs MC-ensemble selections.** Where both are available, report the intersection and differences to justify reliance on ensemble rules when CleanLab cannot run end-to-end.

6.4 Limitations of the statistical analysis

Dependence on the test set. All statistical tests assume the test set is representative. If the test set itself contains label noise, effect-size estimates may be attenuated or biased. We used a stratified split and inspected samples to reduce this risk.

Multiple comparisons. We report many per-class metrics; we did not apply family-wise correction for multiple hypothesis tests on per-class changes (report these as descriptive). The primary inference uses Macro-F1 and the McNemar test on overall paired errors.

Potential for introduced bias. Automatic relabelling can introduce systematic biases if the model's errors correlate with protected attributes or topical skew. Our conservative policy and spot-checks mitigate but do not eliminate this risk; further human auditing is prudent for deployed systems.

6.5 Summary & interpretation

Both McNemar's exact test and bootstrap CI for Δ(Macro-F1) provide strong evidence that the conservative automatic relabelling pipeline yielded a real improvement on the held-out test set. Error analysis shows that gains are concentrated where metadata errors were present and in several mid-frequency classes; remaining failures mostly involve long-document signal outside the truncation window, ambiguous multi-topic pages, or non-article pages. These insights motivate targeted next steps (chunking/long-context modeling for long docs, multi-label treatment for ambiguous pages, and additional doc-type filtering) which we discuss in §7.

7. Discussion

7.1 Practical Implications and Deployment

Our pipeline demonstrates a cost-effective approach to label noise reduction. Using MC-dropout from a single model provides robust uncertainty signals without the cost of multiple models, and truncation enables rapid iteration. For deployment, we recommend offline execution of the noise detection, careful validation of auto-relabels, and maintaining audit logs.

The choice between automated relabelling and manual review depends on the context. Automation is suitable when label errors are concentrated and high-precision corrections are sufficient, as in our study. Manual review is preferable for numerous errors, sensitive categories, or when class definitions are ambiguous.

7.2 Limitations

We identify several limitations that temper the generality of our findings:

Dependence on the teacher model. The quality of candidate detection depends on the teacher's learned priors; a biased teacher can produce high-confidence but incorrect relabel proposals. Conservative thresholds and ensemble signals mitigate but do not eliminate this risk.

Truncation blind spots. Using only the first 512 tokens misses discriminative evidence that appears later in long articles; this limits both teacher accuracy and the reliability of relabel proposals for long-form content. Chunking or long-context models are natural remedies but require more compute and engineering.

Potential for introduced bias. Automated relabelling may reinforce existing dataset biases (class priors, topical skews). We recommend targeted audits for sensitive attributes and class-level relabel counts before wide adoption.

Domain specificity and generalization. Our experiments use an Arabic news/portal crawl; results and decision thresholds may not transfer directly to other domains or languages without retuning.

Evaluation noise. If the held-out test set contains label noise, effect-size estimates may be conservative or biased; we reduced this risk by using stratified splits and manual spot checks, but an entirely clean test set is ideal for the strongest claims.

7.3 Ethical & copyright considerations

Copyright & data sharing. The corpus originates from publicly available web content. We publish only sanitized artifacts, processing scripts, and a small set of anonymized text examples sufficient for basic reproducibility. Any broader distribution of raw text would be subject to the original source's terms of use and applicable copyright law. We record these constraints in our internal documentation and will include a redacted summary of compliance requirements in the artifact bundle provided upon acceptance.

Automated label changes and downstream harms. Automated relabelling may have downstream effects (e.g., on search, recommendation, or moderation systems). We therefore adopt a conservative relabelling policy, keep explicit internal logs of all automatic edits, and recommend a human-in-the-loop review for sensitive categories. Redacted summaries of diagnostic edits and the decision rules used will be provided upon acceptance; full internal logs are retained for audit under appropriate access controls.

Privacy & anonymization. Any released examples are sanitized to remove personally identifying information and other sensitive content. The processing pipeline and any released artifacts are intended to comply with institutional and legal privacy requirements; requests for access to more detailed data or logs will be handled only after appropriate approvals and in accordance with the source terms and copyright constraints.

8. Conclusion and Future Work

8.1 Conclusion

We presented a practical, reproducible pipeline to detect and correct label noise in a large Arabic web-article corpus. Using an AraBERTv02 teacher, MC-dropout ensemble signals, and conservative decision rules inspired by confident learning, we automatically relabeled a small set of training examples (880 in the reduced experiment) and observed a statistically significant improvement on the held-out test set (Macro-F1 from 0.9367 to 0.9503, McNemar $p \approx 1.9\times10^{-5}$, 95% bootstrap CI for Δ = [0.0082, 0.0190]). Our analysis shows that high-precision automatic relabelling can yield meaningful gains with modest compute and without extensive manual annotation, particularly when noisy labels are concentrated in a modest subset of the data.

8.2 Future work

Key directions we plan or recommend:

Scale cleaning to the full corpus. Apply the same conservative pipeline to the full 361k+ training split and re-evaluate gains at full scale.

Hybrid chunking / long-context strategy. For long documents, compare chunk-aggregation strategies against a long-context transformer (e.g., Longformer/BigBird) in terms of both accuracy gains and computational cost; use chunk-level agreement as an extra signal for relabel decisions.

Human-in-the-loop calibration. Collect a small high-quality annotated calibration set (e.g., 1k examples) to tune thresholds for target precision levels and to measure the empirical precision of auto-relabels across classes.

Active learning & semi-supervised extensions. Use active selection to prioritize ambiguous or high-impact examples for manual labeling, and explore semi-supervised training (pseudo-labeling) coupled with label-quality weighting to further improve robustness.

Bias & fairness audits. Perform systematic analyses to detect whether relabelling shifts model behavior across demographic or topical subgroups, and implement safeguards (e.g., constrained relabel rules) if needed.

Generalization to other languages and domains. Validate the pipeline on other web-crawled corpora and languages to assess transferability of thresholds and rules.

References

[1] Beltagy, I., Peters, M. E., Cohan, A.: Longformer: The Long-Document Transformer. arXiv:2004.05150 (2020). https://arxiv.org/abs/2004.05150. (Accessed Nov 15, 2025)

[2] Zaheer, M., et al.: BigBird: Transformers for Longer Sequences. NeurIPS (2020). https://papers.neurips.cc/paper/2020. (Accessed Nov 15, 2025)

[3] Northcutt, C. G., Jiang, L., Chuang, J.: Confident Learning: Estimating Uncertainty in Dataset Labels. arXiv:1911.00068 (2019). https://arxiv.org/abs/1911.00068. (Accessed Nov 15, 2025)

[4] Antoun, W., Baly, F., Hajj, H.: AraBERT: Transformer-based Model for Arabic Language Understanding. In: Proceedings of OSACT (2020). Available: https://aclanthology.org/2020.osact-1.2/. (Accessed Nov 15, 2025)

[5] aubmindlab: bert-base-arabertv02 (model card). Hugging Face Hub. https://huggingface.co/aubmindlab/bert-base-arabertv02. (Accessed Nov 15, 2025)

[6] Gal, Y., Ghahramani, Z.: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In: ICML (2016). https://proceedings.mlr.press/v48/gal16.html. (Accessed Nov 15, 2025)

[7] Sujan Hiregundagal Gopal Rao. (2023). A Review of Cybersecurity Threats in Automotive Semiconductor Control Units. International Journal of Intelligent Systems and Applications in Engineering, 12(1), 927–932.

[8] Principe, R. Alva, et al.: Long Document Classification in the Transformer Era: a review (2024/2025). Wiley / survey (2025). (Accessed Nov 15, 2025)

[9] McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12(2), 153–157 (1947).

[10] cleanlab: Cleanlab open-source library and documentation. GitHub & docs. https://github.com/cleanlab/cleanlab and https://docs.cleanlab.ai/ (Accessed Nov 15, 2025)

[11] Yu, C., Ma, X., Liu, W.: *Delving into Noisy Label Detection with Clean Data.* In: Proceedings of the 40th International Conference on Machine Learning (ICML), PMLR (2023). Full text (conference PDF).

[12] Lin, T., Wang, M., Lin, A., Mai, X., Liang, H., Tham, Y.-C., Chen, H.: *Efficiency and safety of automated label cleaning on multimodal retinal images.* npj Digital Medicine, vol. 8, article 10 (2025). https://www.nature.com/articles/s41746-024-01424-x.

[13] Shelmanov, A., Tsymbalov, E., Puzyrev, D., Fedyanin, K., Panchenko, A., Panov, M.: *How Certain is Your Transformer?* In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021), pp. 1833–1840 (2021). DOI: 10.18653/v1/2021.eacl-main.157.

[14] Daouadi, K. E., Boualleg, Y., Haouaouchi, K. E.: *Ensemble of pre-trained language models and data augmentation for hate speech detection from Arabic tweets.* arXiv:2407.02448 (2024). https://arxiv.org/abs/2407.02448.

[15] Wolf, T., et al.: Transformers: State-of-the-Art Natural Language Processing. EMNLP System Demonstrations (2020). https://huggingface.co/transformers/ (Accessed Nov 15, 2025)

[16] Efron, B., Tibshirani, R. J.: An Introduction to the Bootstrap. Chapman & Hall/CRC (1993).

[17] Yagci, U., Iscan, E., Kolcak, A. E.: *ReBERT at HSD-2Lang 2024: Fine-Tuning BERT with AdamW for Hate Speech Detection in Arabic and Turkish.* In: Proceedings of the 7th CASE Workshop (CASE 2024) (2024). ACL Anthology: https://aclanthology.org/2024.case-1.27/