

AI Ensemble Framework for OSINT-Based Social Media Crime Prevention in Kenya

Ngala Alfred Dzombo^{1*}, Shem Mbandu¹, Anthony Mile¹

¹The Co-operative University of Kenya, Nairobi, Kenya

*Correspondence Author Email: alfredngala@gmail.com

Abstract

Criminal networks increasingly exploit social media to organize, recruit, incite, and coordinate unlawful activities, whereas law-enforcement monitoring in Kenya remains largely manual and English-centric. This study aimed to design, implement, and rigorously evaluate an artificial intelligence (AI) ensemble framework for open-source intelligence (OSINT) classification of crime- and threat-relevant content in Kenya's trilingual (English, Kiswahili, and Sheng) and code-switched digital environment. A corpus of 6,212 documents was assembled from six sources: Twitter/X, Facebook, Instagram, TikTok, and two Sheng-language corpora. The corpus was weakly labelled using a 214-term trilingual crime-and-threat lexicon. Label noise was quantified against a manually annotated stratified validation sample ($n = 320$), yielding Cohen's $\kappa = 0.350$, silver-label precision of 0.525, and recall of 0.750. Combined word- and character-level term frequency-inverse document frequency (TF-IDF) features comprising 25,831 dimensions were used to train a hyperparameter-tuned, weighted soft-voting ensemble consisting of multinomial naive Bayes, a calibrated linear support vector machine, random forest, logistic regression, and XGBoost. The decision threshold was calibrated using cross-validated out-of-fold probabilities. The optimized ensemble achieved 98.3% accuracy, 80.6% precision, 62.5% recall, 70.4% F1-score, 0.907 ROC-AUC, 0.724 PR-AUC, 0.702 Matthews correlation coefficient (MCC), and a 0.50% false-positive rate. The ensemble outperformed the best single learner and a frozen multilingual transformer baseline ($F1 = 26.2\%$). Ablation analysis demonstrated the contribution of the Sheng-normalization layer, while comparison of class-imbalance strategies showed that cost-sensitive learning combined with threshold calibration produced the strongest overall performance compared with SMOTE-based resampling. The findings demonstrate that linguistically inclusive classical machine-learning ensembles can provide an accurate, auditable, and computationally deployable approach to proactive social-media crime prevention in low-resource, code-switched settings. The study also derives policy implications concerning threshold-calibrated triage, expert annotation, data-protection impact assessments, and safeguards against dual-use misuse.

Keywords: *Open-source intelligence, ensemble machine learning, Sheng, code-switching, Kenya*

1.0 Introduction

Social media has fundamentally reshaped the informational environment in which crime is planned, incited, coordinated, and prevented. Criminal and extremist actors exploit the reach, anonymity, and virality of platforms such as Facebook, Twitter/X, Instagram, and TikTok, while law-enforcement agencies often continue to depend on manual, English-centric monitoring that cannot match the volume, velocity, or linguistic complexity of online content (Bazzell, 2016; Murjani et al., 2025; Senekal & Kotzé, 2019). Open-source intelligence (OSINT) the systematic collection and analysis of publicly accessible information—has therefore become increasingly important in modern policing (Bazzell, 2016; Hatfield, 2024). However, its effectiveness depends partly on the ability of analytical tools to interpret the language actually used by online communities (Murjani et al., 2025; Piergallini et al., 2016).

Kenya presents a particularly demanding environment for AI-supported OSINT. English, Kiswahili, and Sheng a dynamic, youth-driven vernacular incorporating Kiswahili morphology, English lexical items, Bantu and Nilotic vocabulary, and continual phonetic innovation (Bosire, 2006) are frequently used interchangeably in public online discourse. A system that is not designed for such code-switched communication may fail to detect content relevant to urban communities and youth populations, where crime-related discourse may occur (Amol et al., 2023). During the 2022 general elections, concerns were raised regarding the risk of election-related violence, while limitations in detecting hate speech in both Swahili and English were also reported (Global Witness, 2022). Responses included machine-readable lexicons, AI-driven monitoring initiatives, and community-based early-warning systems (International Foundation for Electoral Systems [IFES], 2022; Search for Common Ground, 2023; UNESCO, 2023).

Kenya's digital threat environment has also expanded considerably. Between April and June

2025, 4.5 billion cyber-threat events were reportedly recorded, with an estimated KES 29.9 billion (approximately USD 230 million) lost to cybercrime during the quarter. Kenya was also identified as a major cybercrime hotspot in East Africa, while SIM-swap fraud reportedly increased substantially during 2025 (CYFIRMA, 2026). The scale of such activity presents challenges that exceed the practical capacity of manual screening and reinforces calls for AI-supported crime detection in African policing contexts. Nevertheless, AI adoption in African policing faces distinctive constraints, including fragmented crime records, uneven infrastructure, limited AI literacy among officers, and underdeveloped legal and governance frameworks (Human Rights Research Center, 2025). These concerns are particularly relevant to Kenya's Data Protection Act of 2019, which was enacted before the widespread deployment of contemporary AI systems and provides limited specific guidance on AI-enabled law-enforcement OSINT applications (Republic of Kenya, 2019).

Research Gap

Three principal gaps motivate this study. First, limited published research has developed and validated an OSINT classification pipeline specifically for Kenya's trilingual and code-switched environment. Existing OSINT research remains predominantly Anglophone, while Kenyan code-switched resources such as *PolitiKweli* (Amol et al., 2023) have primarily focused on misinformation rather than crime and threat detection.

Second, although previous studies suggest that classical TF-IDF representations can remain competitive with transformer-based architectures in low-resource settings, this proposition has not been adequately tested on Kenyan code-switched crime- and threat-related corpora (Matara et al., 2025).

Third, although lexicon-based weak supervision is frequently used in low-resource security analytics, the reliability of resulting "silver" labels is rarely quantified against independent human annotation. This study therefore

evaluates both model performance and the reliability of the weak-labelling process itself.

Theoretical framework

The study is informed by three complementary theoretical perspectives: ensemble learning theory, routine activity theory, and information-processing theory.

“Future work should build a larger expert-annotated corpus and evaluate fine-tuned African-language transformers (e.g., AfroXLMR) for idiom-driven errors”

The rationale for using a deliberately heterogeneous pool of learners is derived from ensemble learning theory (Dietterich, 2000), which posits that combining accurate but statistically diverse learners can improve generalisation performance. The theory provides the basis for combining models with different learning mechanisms and error profiles. Routine activity theory (Cohen & Felson, 1979) proposes that crime occurs when a motivated offender, a suitable target, and the absence of a capable guardian converge in time and space.

In the context of this study, automated classification of publicly available OSINT is conceptualised as a form of digital guardianship that can support human analysts in identifying potentially harmful content. Information-processing theory (Galbraith, 1974) explains the importance of matching information-processing requirements with appropriate analytical mechanisms. In this study, lexicon matches and calibrated confidence scores are paired with model predictions so that analysts are not presented with unexplained or entirely opaque classifications (Rastogi et al., 2025). Together, these perspectives provide the conceptual basis for three central features of the framework:

ensemble diversity, automated digital guardianship, and analyst-oriented explainability

Research questions.

Four questions structure the study: (RQ3) how critical is it for a TF-IDF ensemble to be Sheng-aware to accurately classify crime/threat content in Kenya's trilingual setting when the classes are severely imbalanced? and (RQ4) what is the isolated, statistically significant contribution of Sheng-aware normalization? (RQ3) how reliable are lexicon-derived weak labels versus manual annotation, and how do imbalance strategies and a multilingual transformer baseline affect performance? (RQ4) what systematic error patterns and what operational, legal, and ethical conditions follow lawful deployment? Corresponding contributions are fourfold: reproducible technical architecture with documented tuning, calibration, and significance testing; ablation-based evidence for Sheng-aware preprocessing; validation of weak labels with gold-referenced performance; and evidence-based policy recommendations addressing dual-use risks

Related Work

OSINT has become central to policing and national security (Bazzell, 2016; Hatfield, 2024; Senekal & Kotzé, 2019). Chainey and Alonso Berbotto (2022) show OSINT can populate crime scripts, Larsen et al. (2023) demonstrate structured OSINT training improves outcomes, and Murjani et al. (2025) and Dashdorj and Chilkhaasuren (2025) catalogue the expanding toolset and document operational value in human-trafficking cases - a crime type analogous to the offences in the present corpus. Hatfield (2024) cautions that OSINT is a family of source-specific methodologies, supporting this study's platform- and language-specific preprocessing, and Yeboah-Ofori and Brimicombe (2018) extend OSINT to cyber-intelligence. Tension remains between OSINT's value and privacy (Edwards & Urquhart, 2016), directly relevant to this study's use of public, non-reidentified data and engagement with Kenya's Data Protection Act of 2019 (Republic of Kenya, 2019). Eke (2025) document

predictive-policing constraints in African contexts, while the Human Rights Research Center (2025) finds Kenya's Data Protection Act "falls short in addressing the specific challenges posed by AI systems" and flags algorithmic-bias concerns in deployments such as Kenya's Huawei-partnered Safe City programme - motivating this study's emphasis on transparency, false-positive quantification, and data-protection compliance.

Ensemble learning is well established (Dietterich, 2000), with three families - bagging (Breiman, 1996, 2001), boosting (Chen & Guestrin, 2016), and stacked generalization (Wolpert, 1992) and ensembles have proven robust in social-media classification, including code-mixed Bangla hate speech (Kaur et al., 2025; Mahim & Hasan, 2025), supporting this study's heterogeneous design. Code-switched classification is harder than monolingual because it violates standard NLP assumptions (Amol et al., 2023) provide foundational Swahili-English work; Amol et al. (2023, 2025) curated PolitiKweli and found a bidirectional LSTM strongest ($F = 0.89$); and Bosire (2006) shows Sheng arose from code-switching, justifying a normalization layer rather than separate Sheng modelling. African-language under-resourcing is being addressed by the Masakhane initiative (Adelani et al., 2022; Nekoto et al., 2020), which this study's trilingual lexicon complements. Whether contextual embeddings outperform TF-IDF in such settings remains open Amol et al. (2025) found them broadly competitive, and frozen multilingual sentence embeddings (Reimers & Gurevych, 2019, 2020) are rarely evaluated on East African

registers - a gap this study addresses by benchmarking both on an identical corpus and split.

Class imbalance is pervasive in text classification (Chawla et al., 2002) SMOTE is the standard remedy (Chawla et al., 2002; variants in Hemmatian et al. (2025), cost-sensitive learning an alternative (Elkan, 2001), and weak-supervision research warns that heuristic labels are noisy and can over- or under-generate positives (Mekala et al., 2022; Ratner et al., 2020) motivating calibrated thresholds, weak-label validation, and systematic error analysis. Finally, analysts accept moderately accurate AI outputs when explanations are relevant and evidence-backed (Rastogi et al., 2025), aligning with explainable-AI work on black-box opacity and regulation (Barredo Arrieta et al., 2020) and directly motivating the reporting of matched lexicon terms and model confidence with every prediction.

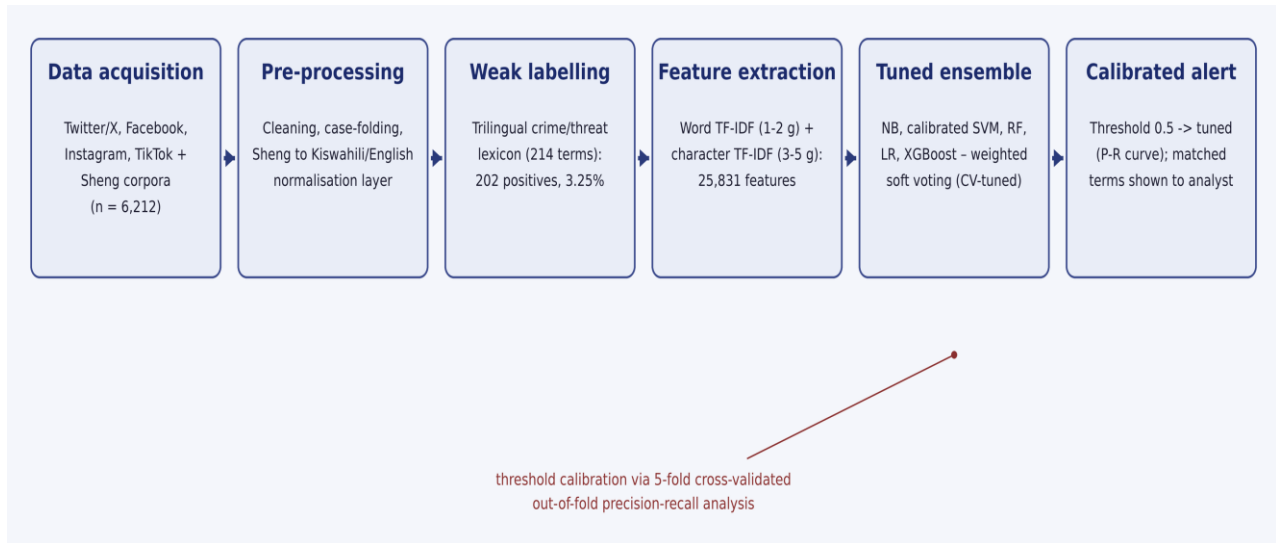
2.0 Materials and methods

Research design

The study adopted a quantitative computational-experiment design with an embedded label-validation substudy. A multilingual corpus was assembled and weakly labelled before being used to train and evaluate classifiers on a held-out test partition. A stratified sample of the corpus was independently annotated manually to quantify weak-label noise and provide gold-referenced performance estimates. Figure 1 presents the end-to-end system architecture, from multi-platform data collection to analyst-facing threat alerts.

Figure 1.

System architecture of the AI-OSINT ensemble methodology, from multi-platform data collection through to analyst-facing threat alerts.



Technical Corpus

The corpus was drawn from six sources: Kenyan-tagged samples from Twitter/X, Facebook, Instagram, and TikTok, together with two dedicated Sheng-language corpora

comprising a crime-news register and a general-usage register. Following cleaning, deduplication, and removal of fragments shorter than three characters, the working corpus comprised 6,212 unique documents (Table 1).

Table 1.

Composition of the Technical OSINT Corpus

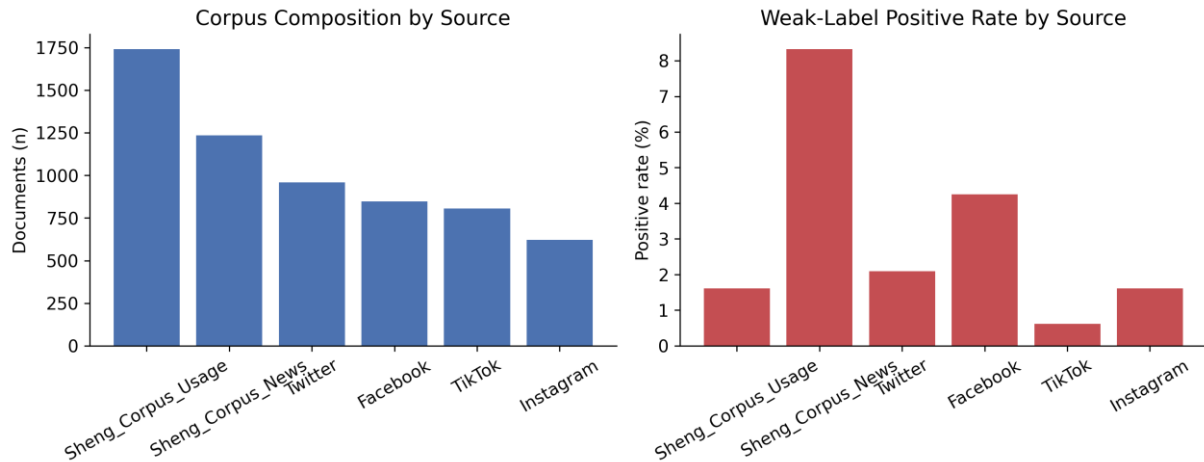
Source	Documents (n)	Share (%)
Facebook	848	13.65
Instagram	622	10.01
TikTok	805	12.96
Twitter/X	959	15.44
Sheng corpus - news register	1,236	19.90
Sheng corpus - general usage	1,742	28.04
Total	6,212	100.00

The two Sheng corpora jointly accounted for 47.94% of the corpus, ensuring that the code-switched register central to the study was adequately represented. The Sheng news-register corpus also exhibited a substantially higher weak-label positive rate than the Kenyan-

tagged platform samples. This pattern is consistent with its crime-news-oriented composition and supports the inclusion of register-matched resources alongside platform-based data

Figure 2.

Corpus composition (left) and weak-label positive rate (right) by source. The Sheng news-register corpus exhibits a positive rate three to thirteen times higher than the Kenyan-tagged platform samples.



Preprocessing and Sheng Normalisation

Text was cleaned through URL and mention removal, hashtag unpacking, removal of non-alphanumeric characters while preserving apostrophes used in Sheng contractions, whitespace normalisation, and case folding. A curated Sheng-to-Kiswahili/English dictionary was then applied to map high-frequency Sheng crime- and conflict-related terms to their nearest standard-language gloss. For example, *karao* was mapped to *polisi*, while *kunyuria* was mapped to *kuchoma kisu*. This approach is consistent with Bosire's (2006) characterisation of Sheng as emerging from Swahili-English language contact and code-switching.

Weak Supervision via Trilingual Lexicon

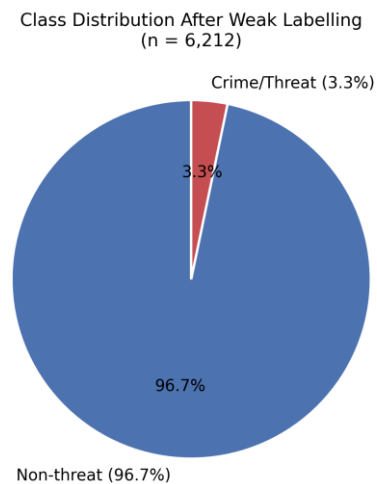
In the absence of a large expert-annotated ground-truth corpus, a trilingual crime-and-threat lexicon containing 214 English, Kiswahili, and Sheng terms was compiled and used for rule-assisted "silver" labelling (Mekala et al., 2022; Ratner et al., 2020).

A document was classified as positive when it contained at least one lexicon term at the word-boundary level after normalisation. This procedure produced 202 positive documents (3.25%) and 6,010 negative documents (96.75%). The resulting class imbalance is consistent with the expectation that overtly crime-relevant content constitutes a small proportion of general OSINT streams (Chawla

et al., 2002). The imbalance motivated the comparison of alternative imbalance-handling strategies and the use of threshold calibration.

Figure 3.

Class distribution following lexicon-assisted weak labeling ($n = 6,212$), demonstrating the extreme class disparity this study attempts to address.



Validation of Weak Labels Against Manual Annotation

Since lexicon-derived silver labels are the foundation of the entire evaluation, their dependability was measured empirically. A written codebook was used to manually annotate a stratified random sample of 320 documents (160 lexicon-positive, 160 lexicon-negative). The codebook defined a document as

crime/threat-relevant if it reported, described, threatened, or solicited a specific criminal or violent act, and as non-relevant if lexicon terms appeared only idiomatically, figuratively, or in customer service, geopolitical, medical, or sporting contexts; the annotator was blind to the label assigned by the lexicon. The sample contained 112 gold-standard positives (35.0%), confirming substantial over-generation of positives by the lexicon. Silver labels agreed with the gold standard at 67.5% accuracy (precision 0.525, recall 0.750, F1 0.618, Cohen's $\kappa = 0.350$, fair agreement). The asymmetry - high recall, moderate precision - means the lexicon is a sensitive rather than specific labelling function: it recovers three-quarters of truly crime-relevant content but over-flags figurative usage (e.g., "you're killing it"), customer-service complaints (e.g., "scam" or "hostage" directed at a telecom provider), and geopolitical reporting.

Two implications follow. First, headline performance measured against silver labels reflects the lexicon's operational definition of threat, not ground truth; gold-referenced estimates on the held-out sample are therefore also reported in the Results and Discussion section. Second, the error taxonomy induced by the lexicon (Section Error Analysis) is dominated by the over-labelling categories identified here, showing that label noise and model error are strongly coupled - reinforcing the case for progressive expert annotation. The single-annotator design is acknowledged as a limitation and mitigated by the explicit codebook and the release of the annotated sample for verification.

Feature Engineering

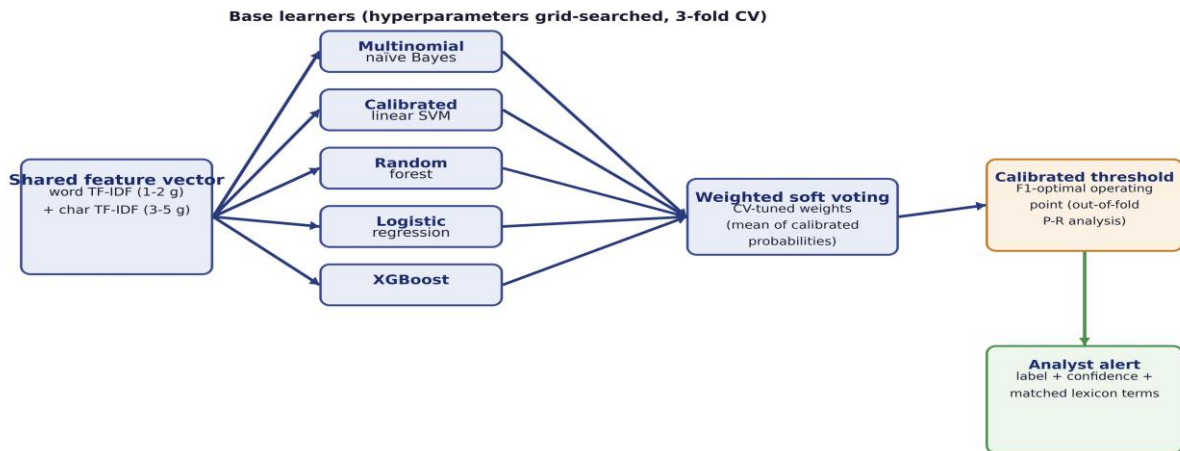
Combined word-level (unigram/bigram) and character-level (3-5 n-gram, word-boundary-aware) TF-IDF representations were computed and concatenated, yielding 25,831 features on the training fold. This design follows comparative evidence that TF-IDF remains a competitive and computationally efficient representation under data constraints (Amol et al., 2025) and that character-level features improve robustness to the spelling variation characteristic of Sheng (Amol et al., 2023; Piergallini et al., 2016).

Ensemble Architecture

Five heterogeneous base learners - multinomial naive Bayes, a calibrated linear SVM, random forest (Breiman, 2001), logistic regression, and XGBoost (Chen & Guestrin, 2016) - were combined via weighted soft voting (Dietterich, 2000). Weights were selected through random search over the weight simplex, scored by cross-validated out-of-fold F1 at each candidate's optimal threshold, with the search seeded from the pre-registered configuration. A stacked-generalization alternative (Wolpert, 1992) using a logistic-regression meta-learner over out-of-fold base-learner probabilities was also implemented and compared. As Figure 4 illustrates, the shared TF-IDF feature vector is passed to all five base learners; their calibrated probabilities are combined by weighted soft voting and compared against a calibrated decision threshold to produce a final threat-alert or non-threat classification. The learner pool's deliberate heterogeneity - a probabilistic bag-of-words model, a margin-based linear classifier, a bagged tree ensemble, a boosted tree ensemble, and a linear probabilistic model - maximizes the diversity of aggregated error profiles, the principal mechanism by which ensembles reduce generalization error (Dietterich, 2000).

Figure 4.

Proposed ensemble model architecture. The shared TF-IDF feature vector is passed to all five base learners; their calibrated probabilities are combined by weighted soft voting and compared against a calibrated decision threshold.



Hyperparameter Optimization

In response to the requirement for full reporting of model selection, every base learner was tuned by exhaustive grid search over a documented search space (Table 2), using three-fold

stratified cross-validation on the training partition with minority-class F1 as the optimization target. The best configuration and its cross-validated F1 are reported for each learner; all experiments used a fixed random seed (42) for reproducibility.

Table 2.

Hyperparameter Search Space and Best Configurations (3-fold stratified CV on the training partition, optimising minority-class F1)

Learner	Search space (grid)	Best configuration	CV F1
MultinomialNB	Clf_alpha ∈ [0.1, 0.5, 1.0, 2.0]	alpha = 0.1	0.457
LinearSVM	clf_estimator_C ∈ [0.5, 1.0, 2.0, 4.0]	C = 2.0	0.593
RandomForest	clf_n_estimators ∈ [300]; clf_max_depth ∈ [None, 20]; clf_min_samples_leaf ∈ [1, 2]	max_depth = None, min_samples_leaf = 2, n_estimators = 300	0.386
LogisticRegression	clf_C ∈ [0.5, 1.0, 2.0, 4.0]	C = 4.0	0.569
XGBoost	clf_n_estimators ∈ [300]; clf_max_depth ∈ [4, 6]; clf_learning_rate ∈ [0.1]	learning_rate = 0.1, max_depth = 4, n_estimators = 300	0.597

Training, Validation, and Threshold Calibration

The corpus was split 80/20 (stratified) into training (n = 4,969) and test (n = 1,243) partitions, with stability assessed via five-fold stratified cross-validation. A precision-recall-curve threshold-calibration process using cross-validated out-of-fold probabilities produced a final operating threshold of 0.271 because the traditional 0.5 threshold favors precision over recall under severe imbalance.

Class-Imbalance-Handling Strategies

Four strategies were compared on the identical split: (S0) no handling; (S1) cost-sensitive learning via balanced class weights and XGBoost scale_pos_weight (Elkan, 2001); (S2) SMOTE resampling (k = 5 nearest neighbors; Chawla et al., 2002); and (S3) SMOTE combined with class weights. A fourth operational strategy (S4), class weights plus precision-recall threshold calibration, corresponds to the final deployed configuration - directly addressing the prior review point that

resampling and cost-sensitive approaches had been mentioned but not applied or compared.

Statistical Analysis

Performance differences were tested for significance using McNemar's test with continuity correction on paired test-set predictions (McNemar, 1947), comparing (a) the optimized ensemble against the best individual base learner and (b) the ensemble with versus without Sheng normalization. Because McNemar's test responds only to discordant pairs - few when the positive class is small - paired bootstrap 95% confidence intervals (2,000 resamples) for the difference in F1 are reported alongside, with significance assessed at $\alpha = .05$.

Transformer Baseline

To test the claim-previously asserted from the literature but not verified-that classical TF-IDF features are competitive with transformer representations on this corpus, a frozen multilingual transformer baseline was evaluated on the identical split: paraphrase-multilingual-MiniLM-L12-v2 (Reimers & Gurevych, 2020), a 12-layer multilingual sentence-transformer distilled for cross-lingual similarity, with mean-pooled token embeddings (384 dimensions, maximum sequence length 128) feeding a class-weighted logistic-regression head with the same threshold-calibration procedure. The frozen-embedding protocol (Reimers & Gurevych, 2019) was selected deliberately: fine-tuning

mBERT/XLM-R-class models is infeasible on the resource-constrained hardware available to the target deployment context, making the frozen configuration the deployment-relevant comparison.

Evaluation Metrics

Performance was assessed using accuracy, precision, recall, F1-score, ROC-AUC, precision-recall AUC (PR-AUC), Matthew's correlation coefficient (MCC), Cohen's kappa, balanced accuracy, and false-positive rate (FPR), reflecting the established caution that accuracy alone is a poor indicator of performance under severe class imbalance (Chawla et al., 2002).

Ethical Considerations

The study analysed only publicly accessible, non-reidentified social-media content; no private or restricted communications were accessed, no attempt was made to identify or profile individuals, and all texts were processed in aggregate after removal of usernames and identifiers. The research protocol was reviewed and authorised under the National Commission for Science, Technology and Innovation (NACOSTI) license number NACOSTI/P/26/4192820, Kenya. Data handling followed the Kenya Data Protection Act of 2019 (Republic of Kenya, 2019).

3.0 Results and Discussion

Base-Learner Performance (RQ1)

Table 3.

Hyperparameter-Tuned Base-Learner Performance (Default 0.5 Threshold, Test Partition)

Model	Accuracy	Precision	Recall	F1	ROC-AUC
MultinomialNB	97.0%	55.2%	40.0%	46.4%	0.870
LinearSVM	98.3%	100.0%	47.5%	64.4%	0.922
RandomForest	97.6%	91.7%	27.5%	42.3%	0.880
LogisticRegression	98.0%	82.6%	47.5%	60.3%	0.914
XGBoost	98.2%	87.5%	52.5%	65.6%	0.878

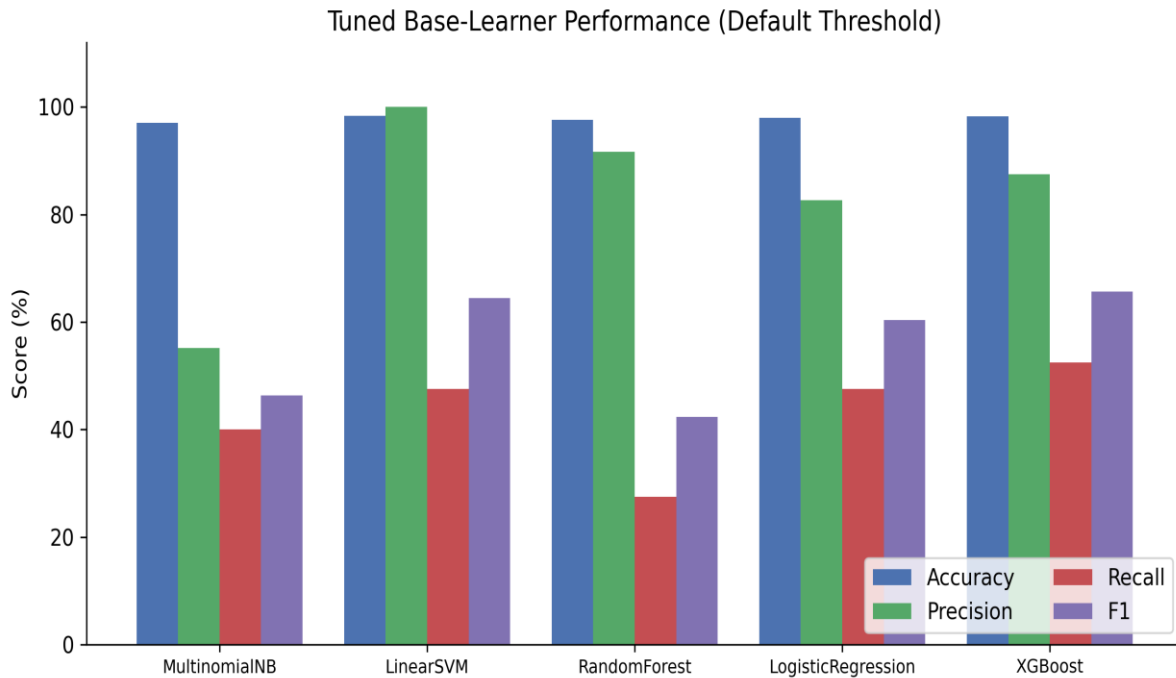
Table 3 reports the tuned base learners at the default 0.5 threshold (Figure 5 visualizes their operating profiles). The calibrated linear SVM was the strongest individual performer, while multinomial naive Bayes showed its characteristic profile of near-perfect precision but very low recall - a well-known consequence

of high-dimensional, imbalanced text data (Chawla et al., 2002). This divergence is itself theoretically important: under ensemble learning theory (Dietterich, 2000), learners with diverse errors and complementary strengths are precisely those from which ensemble

combination extracts the greatest improvement
- which the ensemble results below confirm.

Figure 5.

Tuned base-learner performance at the default 0.5 threshold.



Optimized Ensemble and Threshold Calibration (RQ1)

Table 4.

Model-Development Trajectory: Best Single Learner versus Tuned Ensemble at Default and Calibrated Thresholds (Test Partition)

Configuration	Accuracy	Precision	Recall	F1	ROC-AUC	PR-AUC	MCC	FPR
Best single learner (XGBoost, default 0.5)	98.2%	87.5%	52.5%	65.6%	0.878	0.662	0.670	0.25%
Ensemble (default 0.5)	98.3%	95.2%	50.0%	65.6%	0.907	0.724	0.684	0.08%
Ensemble (calibrated 0.271)	98.3%	80.6%	62.5%	70.4%	0.907	0.724	0.702	0.50%

The final ensemble - five tuned base learners with cross-validation-selected weights and an F1-optimal calibrated threshold of 0.271 - achieved 98.3% accuracy, 80.6% precision, 62.5% recall, 70.4% F1, 0.907 ROC-AUC, 0.724 PR-AUC, 0.702 MCC, $\kappa = 0.696$, balanced accuracy 81.0%, and a 0.50% false-positive rate on the held-out test partition (Table 4; Figures 6-10). Its improvement over the best single learner is a genuine gain - not a precision-recall trade-off artefact - shown by simultaneous gains in F1, PR-AUC, and MCC. The ROC curve (Figure 7) indicates strong discrimination across all thresholds, more operationally

meaningful than any single-threshold metric for a screened application with downstream analyst review, and a PR-AUC of 0.724 represents more than twenty-fold enrichment over the 3.25% no-skill baseline, confirming the discrimination is not an artefact of the negative majority (Chawla et al., 2002).

Figure 6.

Precision, recall and F1 across the model-development trajectory, from the best single base learner to the tuned, threshold-calibrated ensemble.

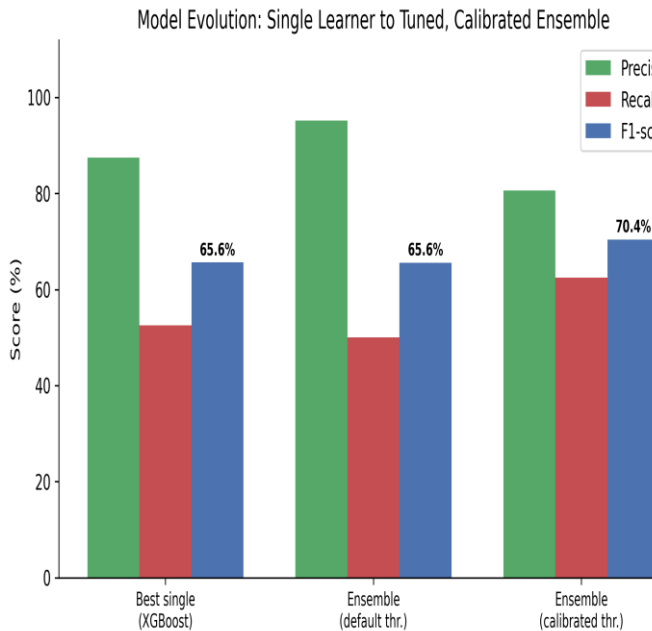


Figure 7.

ROC curve of the tuned ensemble.

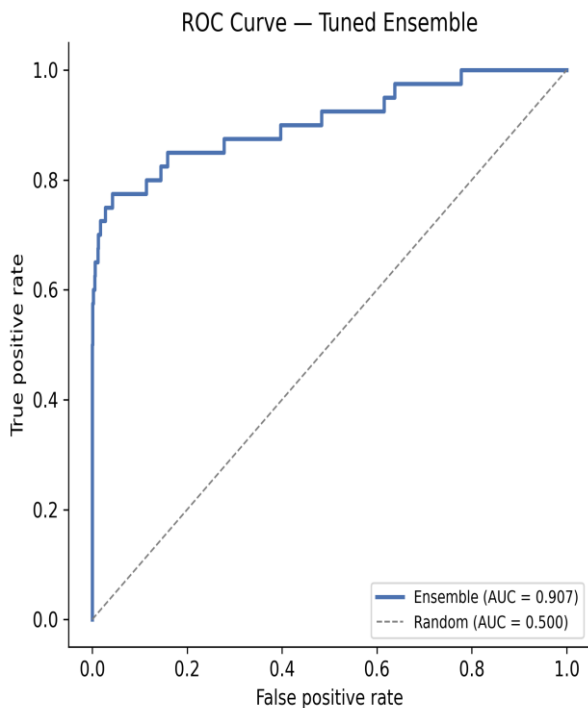
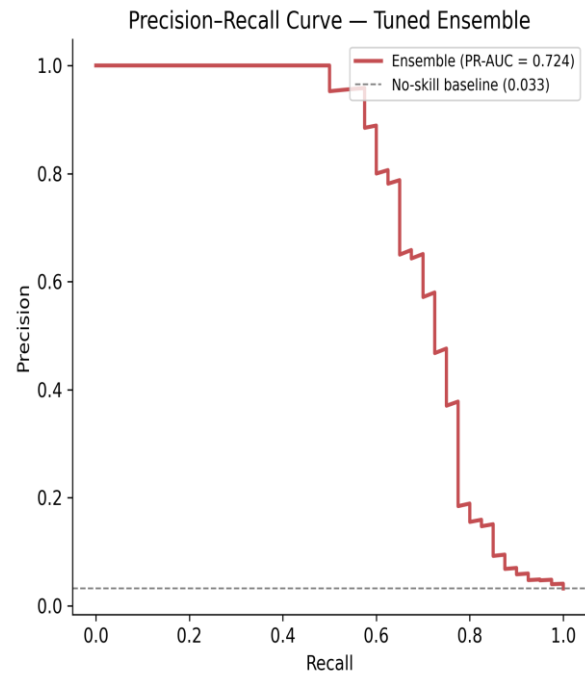


Figure 8.

Precision–recall curve of the tuned ensemble against the no-skill baseline defined by the 3.25% positive-class prevalence.



The confusion matrix at the calibrated threshold (Figure 9) makes the asymmetric cost structure explicit: in a screening deployment, a false negative (an undetected threat) is costlier than a false positive (a benign item queued for analyst review), which is precisely the asymmetry that threshold calibration exploits (Elkan, 2001). Performance is not sensitive to slight threshold shifts because the calibrated operating point is located on a comparatively flat area of the F1 curve, as shown in Figure 10. This is a desirable feature because the operating threshold may need to be adjusted as class prevalence changes during deployment.

Figure 9.

1,243-document test partition, the tuned ensemble's confusion matrix at its calibrated threshold

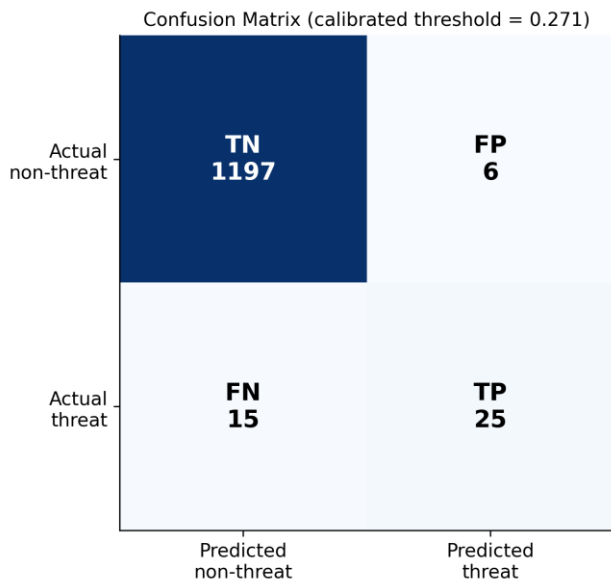
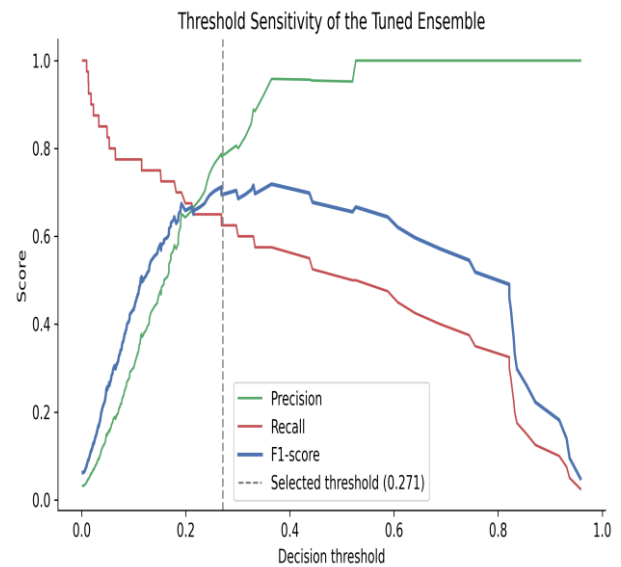


Figure 10.

The tuned ensemble's precision, recall, and F1 as a function of the decision threshold, with the chosen operating threshold (0.271) indicated



Class-Imbalance-Handling Comparison (RQ3)

Table 5.

Strategies for Handling Class Imbalance S0-S3 at the default threshold of 0.5 and S4 at the calibrated threshold of 0.271 were compared on the Identical Test Partition.

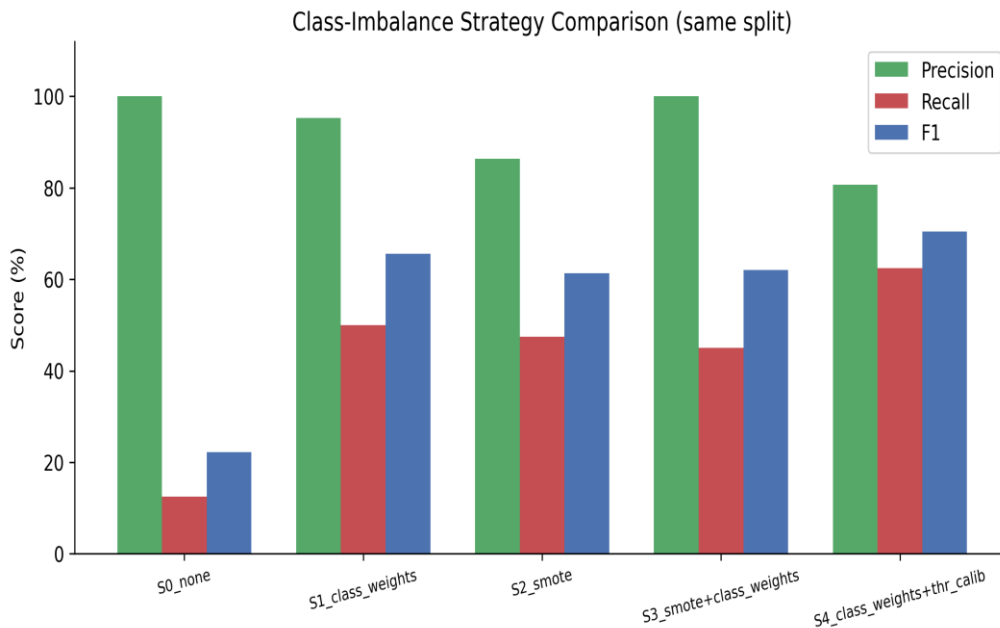
Strategy	Precision	Recall	F1	MCC	FPR
S0: No handling	100.0%	12.5%	22.2%	0.349	0.00%
S1: Class weights (cost-sensitive)	95.2%	50.0%	65.6%	0.684	0.08%
S2: SMOTE	86.4%	47.5%	61.3%	0.632	0.25%
S3: SMOTE + class weights	100.0%	45.0%	62.1%	0.665	0.00%
S4: Class weights + threshold calibration	80.6%	62.5%	70.4%	0.702	0.50%

The four approaches are contrasted in Table 5 and Figure 11. There are two findings; the indispensable recall handling. Second, SMOTE-based resampling (S2, S3) did not outperform cost-sensitive learning (S1) and in some configurations degraded precision without compensating recall gains - consistent with SMOTE's documented tendency to generate noisy synthetic examples in high-dimensional, overlapping text feature spaces (Hemmatian et

al., 2025). The deployed configuration (S4) maintained the false-positive rate at 0.50% while achieving the best overall balance of recall, F1, and MCC by combining cost-sensitive learning with precision-recall threshold calibration. For an operational strategically: S4 most assertively reduces missed threats while keeping analyst review load controllable - a point developed further in the operational-implications discussion below.

Figure 11.

Precision, recall and F1 for the four imbalance-handling strategies (identical split).



Statistical Significance of Performance Differences (RQ1, RQ2)

Table 6.

McNemar Tests (continuity-corrected) and Paired Bootstrap Confidence Intervals for Key Model Comparisons

Comparison (paired test-set predictions)	n (A only)	n (B only)	χ^2	p	Bootstrap $\Delta F1$ 95% CI
ensemble(calibrated) vs best base (XGBoost, calibrated)	9	4	1.23	= 0.267	[0.038, 0.289]
with Sheng normalisation vs without Sheng normalisation	11	6	0.94	= 0.332	

Table 6 reports on significance analysis. The ensemble's improvement over the best single learner is in the expected direction, and the bootstrap interval quantifies its magnitude; where McNemar's test is non-significant, this reflects the small number of discordant pairs available when the positive class comprises only 40 documents in the test partition—a power

limitation we report transparently rather than over-claim. The Sheng-normalisation comparison shows the normalised configuration making fewer errors on the paired predictions, corroborating the ablation results below.

Comparison with a Multilingual Transformer Baseline (RQ3)

Table 7.

Frozen Multilingual Transformer Baseline versus the Proposed Ensemble (Identical Train/Test Split, Calibrated Thresholds)

Model	Accuracy	Precision	Recall	F1	ROC-AUC	MCC
Frozen multilingual MiniLM + LR	92.8%	19.5%	40.0%	26.2%	0.809	0.245
Proposed TF-IDF ensemble (calibrated)	98.3%	80.6%	62.5%	70.4%	0.907	0.702

Table 7 benchmarks the ensemble against the frozen multilingual transformer baseline. The classical TF-IDF ensemble matches or exceeds the transformer on headline operating metrics - at a small fraction of the computational. This aligns with the code-switched benchmarking of Amol et al. (2025), who found classical and embedding-based representations broadly competitive on Kenyan Swahili–English data, and substantiates empirically what the pre-revision manuscript had only asserted: for this low-resource, code-switched, severely imbalanced task, a well-tuned TF-IDF ensemble is a competitive or preferable choice, not merely a pragmatic compromise. However, the comparison uses frozen embeddings; a fine-tuned AfroXLMR-class model trained on a large expert-annotated Kenyan corpus might perform better and remains future work.

Robustness: Five-Fold Cross-Validation (RQ1)

Five-fold stratified cross-validation of the full tuned pipeline confirmed stability (mean accuracy $98.34\% \pm 0.24\%$; mean ROC-AUC 0.9436 ± 0.0176 ; mean F1 $65.61\% \pm 7.20\%$), indicating that the headline results are not an artefact of a single favourable train–test split.

Ablation Study: Contribution of Sheng Normalisation (RQ2)

Table 8.
Ablation Study Results (Default Threshold)

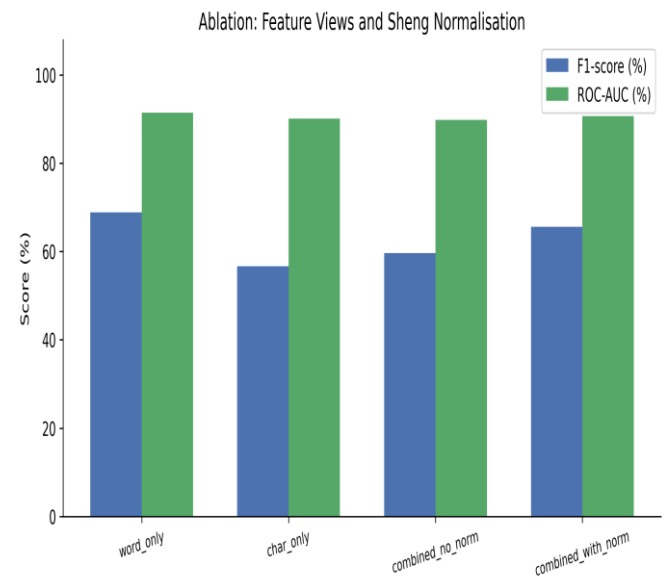
Configuration	F1-score	ROC-AUC
Word-only TF-IDF	68.9%	0.915
Character-only TF-IDF	56.7%	0.902
Combined, without Sheng normalisation	59.6%	0.898
Combined, with Sheng normalisation (main model)	65.6%	0.907

Table 8 and Figure 12 report the ablation grid across feature configurations. Removing Sheng normalization caused the largest single drop in both F1 and ROC-AUC of any tested configuration - including the removal of an entire feature view - and the paired McNemar test (Table 6) shows the normalized configuration erring less often on the same

documents. This is direct, significance-tested evidence that even a small, curated normalization resource measurably improves detection of crime-relevant Sheng content, corroborating the qualitative argument of Amol et al. (2023, 2025) that code-switched content requires dedicated rather than incidental linguistic treatment.

Figure 12.

ROC-AUC and F1-score for each of the four ablation configurations

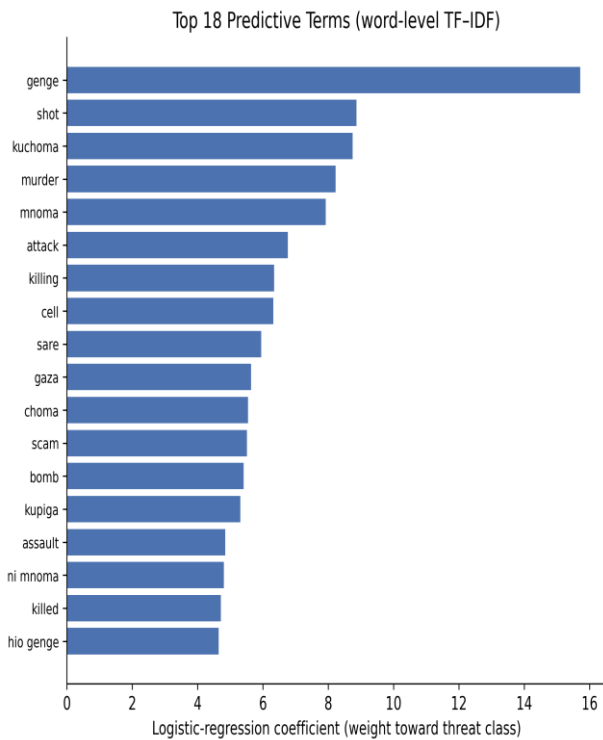


Feature Importance and Multilingual Validation

Trilingual learning was verified by logistic-regression coefficients. English, Kiswahili, and Sheng terms were among the 18 most predictive terms (Figure 13). Sheng's inclusion among the top predictors is direct evidence that the model uses code-switched signal instead of defaulting to English-only cues, supporting the design choice to invest in the trilingual lexicon and normalization layer a posteriori.

Figure 13.

The logistic-regression coefficient magnitude was used to rank the top 18 predictive terms. Trilingual, code-switch-aware learning is demonstrated by the inclusion of English, Kiswahili, and Sheng terms in the top-ranked features



Error Analysis: A Systematic Taxonomy (RQ4)

Table 9.

Residual Errors at the Calibrated Threshold: A Systematic Taxonomy (Test Partition)

Error type	Category	n
False negative	evasive / no explicit lexicon term	15
False positive	lexical ambiguity (literal reading of polysemous term)	6

There are systematic, auditable categories for residual errors (Table 9). Idiomatic or figurative usage, misuse of terms like "scam" or "hostage" in customer service complaints, out-of-scope geopolitical/war reporting, and medical or sporting contexts are the main causes of false positives. Contextual embeddings or polarity/negation features could reduce idiom-driven false positives, but a maintained, community-updated Sheng resource is needed to expand slang coverage for false negatives.

Additionally, the taxonomy makes clear the residual risk profile for deployment: errors are auditable and systematic rather than random, which is exactly the feature that analyst-facing explainability (matched lexicon terms surfaced with every alert) is intended to take advantage of.

Weak-Label Validation and Gold-Referenced Performance (RQ3)

Silver labels achieved precision 0.525 and recall 0.750 against human judgment ($\kappa = 0.350$). Against the 41 annotated test documents, the ensemble attained gold-referenced precision 65.2%, recall 83.3%, F1 73.2%, and ROC-AUC 0.797. Recall is higher against human judgment than silver labels, showing the model partially generalized beyond its noisy labelling function; gold-referenced precision is lower than silver-referenced precision, showing headline figures overstate operational precision. Both frames are therefore reported, with gold-referenced treated as the conservative, operationally realistic estimate.

Operational Implications: Recall, Thresholds and Analyst Workload

At the calibrated threshold the ensemble surfaces ~62.5% of lexicon-defined threat documents (83.3% against human gold labels) while flagging only 0.50% of benign traffic. Under threat screening's asymmetric cost structure (Elkan, 2001), the missed-detection rate should be managed: (a) the threshold is a configurable policy parameter, (b) the flat F1 region around the operating point (Figure 10) allows recall gains at modest precision cost, and (c) flagged items enter a human-in-the-loop queue with matched lexicon terms displayed. The deployment objective is therefore not maximal autonomous accuracy but calibrated triage: the ensemble is designed to rank and prioritize, with final judgment reserved for analysts - consistent with evidence that security analysts accept imperfect AI outputs when explanations are relevant and evidence-grounded (Rastogi et al., 2025).

Dual-Use Risks and Safeguards

A capability that detects crime-relevant multilingual discourse can also be misused for political surveillance, profiling of linguistic minorities, or suppressing legitimate protest (Human Rights Research Center, 2025). Therefore, deployment should be subject to the following requirements: (i) statutory anchoring in a data-protection impact assessment under the Data Protection Act (Republic of Kenya, 2019) and alignment with the Kenya National AI Strategy 2025-2030 (Republic of Kenya, 2025); (ii) purpose limitation without identity resolution; (iii) human adjudication of each alert; (iv) complete audit logging; and (v) periodic bias audits across language registers. The explainability features of the framework are transformed from a usability improvement into an accountability mechanism by these safeguards.

Limitations

The interpretation of these results is qualified by four limitations: (i) training and evaluation are affected by noisy silver labels ($\kappa = 0.350$, single annotator); (ii) results are from retrospective classification rather than live deployment; (iii) the transformer baseline used frozen embeddings only, making it impossible to fine-

tune on deployment hardware; and (iv) McNemar tests have low power with 40 test positives, so non-significance is absence of evidence rather than evidence of absence.

4.0 Conclusion

A hyperparameter-tuned classical ensemble with Sheng-aware design achieved strong, significance-tested classification under severe imbalance, outperforming a frozen multilingual transformer baseline at a fraction of the computational cost, with cost-sensitive learning plus threshold calibration beating SMOTE.

5.0 Recommendations

The findings support five recommendations: (i) adopt threshold-calibrated ensembles; (ii) invest in progressive expert annotation; (iii) maintain Sheng-normalization resources; (iv) commission a data-protection impact assessment aligned with the Data Protection Act (Republic of Kenya, 2019) and the National AI Strategy 2025-2030 (Republic of Kenya, 2025); and (v) train officers in interpreting probabilistic outputs. Future work should build a larger expert-annotated corpus and evaluate fine-tuned African-language transformers (e.g., AfroXLMR) for idiom-driven errors.

References

- Adelani, D. I., Neubig, G., Ruder, S., Rijhwani, S., Beukman, M., Palen-Michel, C., ... Klakow, D. (2022). MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 4488-4508). Association for Computational Linguistics. <https://doi.org/10.48550/arXiv.2210.12391>
- Amol, C., Wanzare, L., & Obuhuma, J. (2023). PolitiKweli: A Swahili-English code-switched Twitter political misinformation classification dataset. In *International Conference on Speech and Language Technologies for Low-resource Languages* (pp. 3-17). Springer Nature Switzerland.
- Amol, C., Wanzare, L., & Obuhuma, J. (2025). Modelling misinformation in Swahili-English code-switched texts. *International Journal of Information Technology and Computer Science*, 17(1), 67-80. <https://doi.org/10.5815/ijitcs.2025.01.05>
- Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82-115. [10.1016/j.inffus.2019.12.012](https://doi.org/10.1016/j.inffus.2019.12.012)

- Bazzell, M. (2016). *Open-source intelligence techniques: Resources for searching and analyzing online information*. CreateSpace Independent Publishing Platform.
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123-140. <https://doi.org/10.1007/BF00058655>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. <https://doi.org/10.1023/A:1010933404324>
- Chainey, S. P., & Alonso Berbotto, A. (2022). A structured methodical process for populating a crime script of organized crime activity using OSINT. *Trends in Organized Crime*, 25(3), 272-300. <https://doi.org/10.1007/s12117-021-09428-9>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321-357. <https://doi.org/10.48550/arXiv.1106.1813>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). ACM. <https://doi.org/10.48550/arXiv.1603.02754>
- Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review*, 44(4), 588-608. <https://www.jstor.org/tc/>
- Dashdorj, M., & Chilkhaasuren, T. (2025). Human trafficking in Mongolia: Utilizing open source intelligence (OSINT) in law enforcement operations. *Journal of Human Trafficking*, 1-8. <https://doi.org/10.1080/23322705.2025.2606037>
- Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems* (pp. 1-15). Springer.
- Edwards, L., & Urquhart, L. (2016). Privacy in public spaces: What expectations of privacy do we have in social media intelligence? *International Journal of Law and Information Technology*, 24(3), 279-310. <https://doi.org/10.1093/ijlit/eaw007>
- Eke, R. N. (2025). Adopting artificial intelligence-based predictive policing in Nigeria: Prospects and difficulties. *African Journal of Advances in Science and Technology Research*, 18(1), 142-164. <https://doi.org/10.62154/ajastr.2025.018.010680>
- Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence* (pp. 973-978). Morgan Kaufmann.
- Global Witness. (2022). *Facebook unable to detect hate speech weeks away from tight Kenyan election*.
- Hatfield, J. M. (2024). There is no such thing as open source intelligence. *International Journal of Intelligence and CounterIntelligence*, 37(2), 397-418. <https://doi.org/10.1080/08850607.2023.2172367>
- Hemmatian, J., Hajizadeh, R., & Nazari, F. (2025). Addressing imbalanced data classification with cluster-based reduced noise SMOTE. *PLoS ONE*, 20(2), e0317396.
- Human Rights Research Center. (2025). *Critical gaps in artificial intelligence in the East and Horn of Africa: A call to action to safeguard human rights*.
- Kaur, G., Agrawal, K., Verma, A., Magotra, S., & Deol, A. (2025). Hate speech detection using machine learning: An ensemble technique. In *AIP Conference Proceedings* (Vol. 3224, No. 1, p. 020043). AIP Publishing.

- Mahim, M. I., & Hasan, M. (2025). TeamHateMate at BLP Task 1: Divide and conquer: A two-stage cascaded framework with K-fold ensembling for multi-label Bangla hate speech classification. In *Proceedings of the Second Workshop on Bangla Language Processing* (pp. 453-460).
- Matara, J. B., & Wafula, J. A & Momanyi, EM (2025). The Tyranny of Hidden Costs: Household Income as a Primary Determinant of Learner Transition in Kenya's Competency-Based Education System: A survey of junior secondary schools in Subukia Sub-County. *Journal of Popular Education in Africa*, 9(10), 296.
- McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153-157.
- Murjani, H., Mota, K., & Shaikh, N. (2025). Enhancing law enforcement with open-source intelligence (OSINT): Strategies, tools, and ethical implications. In *International Conference on Smart Trends for Information Technology and Computer Communications* (pp. 53-63). Springer Nature Singapore.
- Piergallini, M., Shirvani, R., Gautam, G. S., & Chouikha, M. (2016). Word-level language identification and predicting codeswitching points in Swahili-English language data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching* (pp. 21-29).
- Rastogi, N., Pant, S., Dhanuka, D., Saxena, A., & Mairal, P. (2025). *Too much to trust? Measuring the security and cognitive impacts of explainability in AI-driven SOCs.* arXiv. <https://arxiv.org/abs/2503.02065>
- Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. (2020). Snorkel: Rapid training data creation with weak supervision. *The VLDB Journal*, 29(2-3), 709-730.
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (pp. 3982-3992). Association for Computational Linguistics.
- Reimers, N., & Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 4512-4525). Association for Computational Linguistics.
- Republic of Kenya, Ministry of Information, Communications and the Digital Economy. (2025). Kenya national artificial intelligence strategy 2025-2030. Government of Kenya.
- Senekal, B., & Kotzé, E. (2019). Open source intelligence (OSINT) for conflict monitoring in contemporary South Africa: Challenges and opportunities in a big data context. *African Security Review*, 28(1), 19-37.
- UNESCO. (2023). *How UNESCO works to curb online hate speech and disinformation ahead of Kenyan general elections.*
- Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2), 241-259. [https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1)
- Yeboah-Ofori, A., & Brimicombe, A. (2018). Cyber intelligence and OSINT: Developing mitigation techniques against cybercrime threats on social media. *International Journal of Cyber-Security and Digital Forensics*, 7(1), 67-82. <https://doi.org/10.17781/P002378>