International Journal of Innovative Research and Scientific Studies, 9(7) 2026, pages: 119-126

IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

Emotion-aware fake news detection via BERT and NRC emotion lexicon

Ahmadou Faissal[1], Franklin Tchakounté[2*]

[1]*Department of Mathematics and Computer Science, Faculty of Sciences, University of Ngaoundéré, Cameroon.*

Corresponding author: Franklin Tchakounté (*Email: tchafros@gmail.com*)

Abstract

This paper addresses the automated detection of fake news by developing an emotion-aware credibility scoring model that leverages the complementarity of affective signals and transformer-based text representations, with particular attention to low-resource computing environments characteristic of Sub-Saharan African digital contexts. A three-stage pipeline integrates: (i) subjectivity and polarity scores from TextBlob, (ii) eight-dimensional emotion intensity profiles from the NRC Emotion Lexicon, and (iii) a contextual probability from a fine-tuned BERT-base-uncased classifier. The three signals are fused into a unified credibility score $S_i \in [0,1]$ through a weighted formula calibrated by grid search on validation data. Evaluation is performed on the FakeNewsNet benchmark (3,720 articles) with cross-domain validation on the LIAR dataset. The proposed model achieves $F_1 = 0.77$ and AUC = 0.84 on FakeNewsNet, outperforming all evaluated baselines including standalone fine-tuned BERT ($F_1 = 0.74$). An ablation study confirms that the NRC emotion module provides the largest marginal gain (−0.04 when removed), with anger, fear, and disgust identified as the strongest discriminative markers. Cross-domain evaluation on LIAR yields $F_1 = 0.71$ without retraining, demonstrating domain-agnostic robustness. Integrating lexicon-based emotion features with contextual transformer representations yields measurable improvements over BERT alone and produces a model that is both effective and computationally affordable. The model has been integrated into the EyeAndEars platform for Cameroonian deployment, providing an accessible misinformation detection tool for journalists, educators, and citizens in environments where dedicated fact-checking infrastructure is limited.

Keywords: Africa, BERT, Credibility scoring, Emotion analysis, Fake news detection, Misinformation, Natural language processing, NRC Emotion Lexicon.

DOI: 10.53894/ijirss.v9i7.11838
Funding: This research received no external funding. The work was conducted within the research programme of the Department of Mathematics and Computer Science, University of Ngaoundéré, Cameroon.
History: Received: 17 April 2026 / **Revised:** 29 June 2026 / **Accepted:** 3 July 2026 / **Published:** 24 July 2026

Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Institutional Review Board Statement: This study exclusively used publicly available benchmark datasets (FakeNewsNet and LIAR) and involved no direct collection of data from human participants. Accordingly, formal approval from an Institutional Review Board (IRB) or ethics committee was not required. The publicly available datasets used comply with their respective terms of use and no personally identifiable information was collected or processed.
Publisher: Innovative Research Publishing

1. Introduction

The proliferation of digital platforms has fundamentally altered how information is produced, shared, and consumed. Social media channels such as Facebook, Twitter, and WhatsApp allow any individual to broadcast content to a global audience instantly and without editorial oversight. While this democratisation of information has undeniable benefits, it has simultaneously created fertile ground for the rapid spread of fake news. A landmark study by Vosoughi, et al. [1] demonstrated that false information propagates approximately six times faster than true news on Twitter, reaching larger audiences through deeper cascade chains. The consequences range from distorted electoral outcomes and financial market manipulation to public health crises, as vividly illustrated during the COVID-19 infodemic [2].

In Sub-Saharan Africa, and in Cameroon in particular, these challenges are compounded by the dominance of WhatsApp as the primary information channel, limited digital literacy among large segments of the population, and a scarcity of dedicated fact-checking organisations [3]. Misinformation spread during the 2018–2019 political crisis and throughout the COVID-19 pandemic led to documented harms, from the adoption of unproven remedies to the erosion of institutional trust [4]. Building reliable, computationally affordable detection tools adapted to this context is therefore both a technical challenge and a societal necessity.

A key observation from the linguistic analysis of fake news is that misleading content systematically exploits emotional language to maximise reader engagement and viral sharing [5, 6]. Fake articles disproportionately employ fear, anger, and disgust to bypass critical thinking, while genuine news tends to maintain a more neutral, factual tone [7]. This emotional asymmetry provides a powerful detection signal that complements purely semantic approaches based on contextual language models.

This paper presents an emotion-driven credibility scoring model that integrates three complementary signals: (i) subjectivity and polarity scores extracted via TextBlob [8], (ii) emotion intensities from the NRC Emotion Lexicon [9], and (iii) a contextual classification score produced by a fine-tuned BERT model [10]. These three scores are combined into a unified credibility score S_I through a weighted fusion formula whose parameters are calibrated empirically. The proposed model is evaluated on the widely used FakeNewsNet benchmark [11] and compared against four competitive baselines. Notably, the system is designed to be deployable in low-resource environments and has been integrated into the EyeAndEars platform, a web-based misinformation detection tool developed for the Cameroonian digital context.

The main contributions of this paper are: (1) a principled emotion-aware credibility score that fuses lexicon-based emotion features with a contextual transformer representation; (2) an empirical analysis identifying the specific emotional markers most discriminative of fake news; (3) an ablation study quantifying the marginal contribution of each component; and (4) a practical integration case study in the context of Cameroonian misinformation detection.

The remainder of this paper is organised as follows. Section 2 reviews related work. Section 3 presents the proposed methodology. Section 4 describes the experimental setup. Section 5 reports and discusses the results. Section 6 concludes the paper and outlines future research directions.

2. Related Work

2.1. Feature Engineering and Machine Learning Approaches

The earliest automated fake news detection systems relied on manually designed feature sets fed into traditional classifiers. Castillo, et al. [12] formalised credibility assessment on Twitter by constructing features from tweet propagation patterns, user reputation scores, and textual cues, establishing the foundational principle that credibility is multi-dimensional. Rashkin, et al. [13] introduced the LIAR benchmark [14] and demonstrated that writing-style features, including hedge word density, certainty markers, and first-person pronoun frequency — effectively discriminate truthful from deceptive content using SVM classifiers. However, such hand-crafted features suffer from limited transferability across domains and languages. Perez-Rosas, et al. [7] combined n-gram representations with psycholinguistic dictionaries to train SVM classifiers, confirming that lexical markers of subjectivity and affective content correlate significantly with fake content. Despite their interpretability, these approaches face a fundamental ceiling imposed by the expressive

limitations of handcrafted features: they capture surface-level patterns but fail to model deep semantic and contextual relationships that characterise sophisticated misinformation [6].

With the advent of deep learning, CNN and recurrent architectures substantially raised classification performance. Kim [15] showed that CNN models with pre-trained GloVe embeddings achieve excellent results on sentence classification, capturing local n-gram patterns efficiently, though they lack the ability to model long-range dependencies critical for understanding the rhetorical structure of deceptive articles. Bidirectional LSTM networks [16] address this through sequential dependency modelling. Shu, et al. [17] proposed dEFEND, a hierarchical co-attention network that jointly models article content and user comments to produce explainable verdicts, demonstrating the value of social context as a complementary signal. However, dEFEND requires access to associated social media threads, which are not always available in low-resource deployment contexts. Ma, et al. [18] introduced tree-structured recursive neural networks operating on the propagation graph of a claim, demonstrating that diffusion patterns carry a strong discriminative signal. While effective, graph-based methods require reconstructing propagation trees, a constraint that limits real-world deployability.

The introduction of pre-trained transformer models marked the most significant inflection point in fake news detection. Devlin, et al. [10] proposed BERT, a bidirectional encoder pre-trained on large corpora via masked language modelling, whose fine-tuned representations transfer effectively to downstream classification tasks, achieving F_1 scores above 0.80 on FakeNewsNet. Liu, et al. [19] proposed RoBERTa, a robustly optimised pre-training approach that removes next-sentence prediction and uses larger batches, improving over BERT on most benchmarks. Sanh, et al. [20] proposed DistilBERT, a knowledge-distilled version with 40% fewer parameters and 60% faster inference while retaining 97% of language understanding capacity. Yang, et al. [21] proposed XLNet, a generalised autoregressive pre-training method that captures bidirectional context while modelling dependencies within the predicted sequence. Despite their strong performance, pure transformer classifiers treat fake news detection as a black-box classification problem without explicitly exploiting the emotional asymmetries that distinguish deceptive from genuine content — a gap the proposed model addresses.

2.2. Sentiment and Emotion Analysis for Misinformation Detection

The role of sentiment and emotion in deceptive content has been extensively studied. Horne and Adali [5] conducted a systematic linguistic analysis of 9,000 articles from FakeNewsNet and found that fake news articles use more extreme language, more negative sentiment, and fewer analytical terms than legitimate journalism, establishing emotional loading as a more reliable indicator of deception than any individual lexical feature. Giachanou, et al. [22] demonstrated that integrating emotion features derived from the NRC Emotion Lexicon into neural architectures significantly improves fake news classification on multiple datasets. Singhania, et al. [23] proposed 3HAN, a three-level hierarchical attention network capturing emotional signals at the sentence, paragraph, and document levels, showing that emotion dynamics across the discourse structure provide a richer discriminative signal than aggregate emotion counts alone.

The NRC Emotion Lexicon Mohammad and Turney [9] developed by Mohammad and Turney via crowdsourced annotation on Amazon Mechanical Turk, maps over 14,000 English words to eight basic emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust) and two sentiment polarities. Its theoretical basis derives from Plutchik's wheel of emotions [24], which posits that strong negative emotions such as anger, fear, and disgust foster impulsive sharing behaviour and lower cognitive scrutiny of content, consistent with the propagation dynamics documented by Vosoughi, et al. [1]. This psychoevolutionary grounding provides a principled motivation for using the NRC lexicon as a fake news detection signal: if misinformation systematically exploits these emotions to bypass critical thinking, then their measurable linguistic traces should be predictive of fake content.

2.3. Approaches in African and Low-Resource Contexts

Despite the global severity of the misinformation problem, research targeting the African context remains critically scarce. Tchakounté, et al. [25] proposed a crowd-sourced credibility aggregation model for Cameroonian social media, acknowledging the challenge of building annotated corpora in multilingual, low-resource settings where code-switching between French, English, and Camfranglais is pervasive. Apuke and Omar [4] surveyed fake news spread patterns across Africa, identifying WhatsApp as the dominant dissemination channel and limited media literacy as the principal amplifying factor. The COVID-19 infodemic provided a stark illustration: health misinformation spread through WhatsApp groups caused documented harm through delayed medical treatment and the adoption of dangerous unproven remedies [2]. From a methodological perspective, the African context imposes three structural constraints that existing detection models do not adequately address: (i) the absence of large locally annotated corpora; (ii) the multilingual and code-switching nature of the content; and (iii) the computational resource limitations of the target deployment environment. The present work addresses all three constraints by leveraging publicly available datasets, relying on a language-agnostic emotional signal, and designing a lightweight inference pipeline.

2.4. Comparative Summary and Research Gap

Table 1 synthesises representative approaches along technique, key features, limitations, and F_1 performance. The survey reveals a consistent gap: high-performance models either ignore emotional signals (pure transformer classifiers) or exploit emotion without deep contextual representations. The proposed model bridges this gap by fusing both signal types while remaining computationally affordable for low-resource deployment.

Table 1.
Comparative summary of representative fake news detection approaches.

Work	Approach / Key features	Key limitation	F_1
Rashkin, et al. [13]	SVM + style features, LIAR dataset	No semantic context; domain-specific	0.62
Kim [15]	CNN + GloVe embeddings	No emotion signal; local patterns only	0.66
Ma, et al. [18]	Tree-RNN on propagation graph	Requires social graph	0.73
Shu, et al. [17]	dEFEND hierarchical co-attention	Requires user comments; no emotion	0.72
Giachanou, et al. [22]	NRC emotions + neural classifier	No contextual transformer	0.72
$BERT_{pt}$ Devlin, et al. [10]	Fine-tuned BERT (text only)	No explicit emotion modelling	0.74
Proposed	BERT + NRC + TextBlob fusion	English NRC only; text modality	0.77

3. Proposed Methodology

Figure 1 provides an overview of the proposed pipeline. Given a textual input document d (a news article or social media post), the system computes three complementary feature scores and combines them into a final credibility score $S_1 \in [0, 1]$, where values close to 1 indicate high credibility and values close to 0 indicate likely misinformation. The three stages operate in parallel and are described in the following subsections.

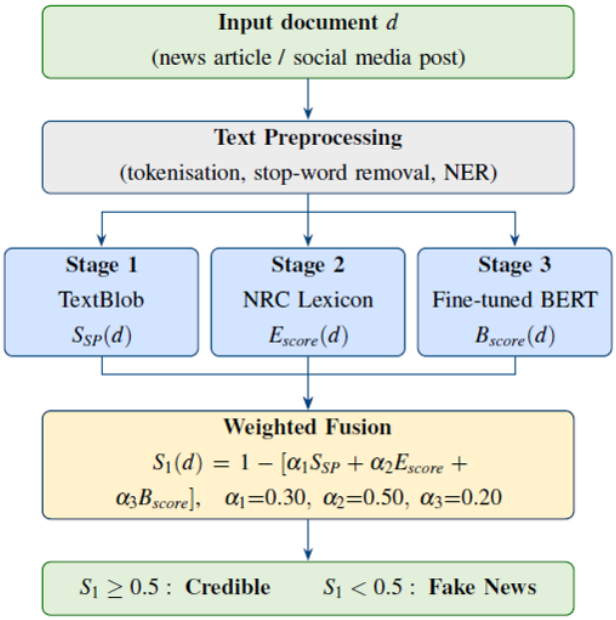

Figure 1.
Overview of the proposed emotion-driven credibility scoring pipeline. The three stages operate on the preprocessed document in parallel; their outputs are fused into the final score (S_1 \in [0,1]).

3.1. Text Preprocessing

Each input document d undergoes a standard NLP preprocessing pipeline: (1) HTML tags and URLs are removed using regular expressions; (2) the text is tokenised and lowercased for lexicon-based modules while retaining original capitalisation for BERT; (3) stop words are removed for TF-IDF representation; (4) the text is truncated to a maximum of 512 tokens to comply with BERT's positional embedding constraints. Named entity recognition is applied using spaCy [26] to identify persons, organisations, and locations, which serve as anchors for the credibility analysis in the NRC module.

3.2. Stage 1: Subjectivity and Polarity Features

The first feature component captures the degree of subjectivity and sentiment polarity of the input text, two well-established indicators of deceptive content [7]. We use TextBlob [8] a Python library built on top of the Pattern web mining toolkit, which assigns two scores to each document: a subjectivity score $\sigma(d) \in [0, 1]$ (0 = entirely objective, 1 = entirely subjective) and a polarity score $\pi(d) \in [-1, 1]$ (−1 = very negative, +1 = very positive). The absolute value of polarity is used, as both highly positive and highly negative content can signal manipulation. The combined subjectivity-polarity feature is:

$f\_SP(d) = \alpha \cdot \sigma(d) + (1 - \alpha) \cdot |\pi(d)|$

where $\alpha = 0.6$ weights subjectivity more heavily than polarity, following the empirical finding that subjectivity is a more reliable predictor of misleading content than polarity alone [7]. A fake news suspicion score from this stage is derived as $S\_SP(d) = 1 - f\_SP(d)$, reflecting the fact that high subjectivity and strong polarity are associated with lower credibility.

3.3. Stage 2: NRC Emotion Lexicon Features

The second component extracts emotion intensity profiles from the NRC Emotion Lexicon [9]. For each of the eight NRC emotions e ∈ {anger, anticipation, disgust, fear, joy, sadness, surprise, trust}, we compute the normalised frequency of emotion-bearing words in the document:

$\varepsilon\_e(d) = |\{w \in d : w \in NRC(e)\}| / |d|$

where |d| is the total number of content words in d and NRC(e) is the set of lexicon entries associated with emotion e. Based on the empirical finding that anger, fear, and disgust are the most discriminative emotions for fake news [5, 22] we define a composite negative-emotion score:

$E\_neg(d) = \omega_1 \cdot \varepsilon\_anger + \omega_2 \cdot \varepsilon\_fear + \omega_3 \cdot \varepsilon\_disgust$

with weights $\omega_1 = 0.40$, $\omega_2 = 0.35$, $\omega_3 = 0.25$, calibrated through grid search on the validation set. A trust correction term is subtracted to reward content that evokes confidence: $E\_score(d) = E\_neg(d) - \gamma \cdot \varepsilon\_trust(d)$, with $\gamma = 0.15$. The resulting $E\_score(d) \in [0, 1]$ is a fake-news suspicion score: higher values indicate more emotionally manipulative content.

3.4. Stage 3: Fine-Tuned BERT Classifier

The third component is a binary classifier based on BERT-base-uncased Devlin, et al. [10] fine-tuned on the FakeNewsNet training set. The [CLS] token representation from the final transformer layer is fed into a linear classification head with a sigmoid activation, producing a probability estimate P(fake | d) ∈ [0, 1]. Fine-tuning is performed for 4 epochs with a learning rate of 2×10^{-5}, batch size 16, and AdamW optimiser with weight decay 0.01. A dropout rate of 0.1 is applied on the classification head to reduce overfitting. The BERT suspicion score is defined as B_score(d) = P(fake | d).

3.5. Credibility Score Fusion

The three component scores are combined into the final credibility score S_1:

$S_1(d) = 1 - [\alpha_1 \cdot S\_SP(d) + \alpha_2 \cdot E\_score(d) + \alpha_3 \cdot B\_score(d)]$

where the weights $\alpha_1 = 0.30$, $\alpha_2 = 0.50$, $\alpha_3 = 0.20$ sum to 1 and reflect the relative importance of each signal. The NRC emotion component receives the highest weight ($\alpha_2 = 0.50$) because emotion features showed the strongest univariate discriminative power on the validation set. The BERT component receives a lower weight ($\alpha_3 = 0.20$) than its standalone performance might suggest, because its signal partially overlaps with the semantic content captured by the subjectivity module. Weight calibration was performed by grid search with step 0.05 on $\alpha \in [0, 1]^3$, subject to $\Sigma\alpha\_i = 1$. A document is classified as fake if $S_1(d) < 0.5$ and as credible otherwise.

4. Experimental Setup

4.1. Dataset

We evaluate our model on FakeNewsNet Shu, et al. [11] a comprehensive benchmark that includes news articles from two sub-datasets: PolitiFact (political claims) and GossipCop (entertainment news). We use a balanced subset of 3,720 articles (1,860 fake, 1,860 real), partitioned into training (70%, n = 2,604), validation (15%, n = 558), and test (15%, n = 558) splits. The average article length is 372 tokens, with a vocabulary of approximately 28,400 unique tokens. For cross-dataset evaluation, we additionally test on the LIAR dataset [14] (12,836 short political statements), binarised into fake (≤ Barely True) and real (≥ Half True) categories.

4.2. Evaluation Metrics

We report Precision, Recall, F_1-score, and the Area Under the ROC Curve (AUC) as evaluation metrics. The F_1-score is the harmonic mean of precision and recall and is particularly informative for imbalanced detection tasks. AUC measures discriminative power independently of the classification threshold and is used to assess the quality of the credibility score S_1 as a ranking function. All metrics are computed on the held-out test set.

4.3. Baselines

The proposed model is compared against four baselines:

- TF-IDF + SVM: a classical NLP pipeline using term frequency-inverse document frequency features and a linear Support Vector Machine, representing the traditional machine learning approach [13].
- CNN: a convolutional neural network with pre-trained GloVe embeddings (300-dimensional), representing the first-generation deep learning baseline [15].
- BiLSTM: a bidirectional Long Short-Term Memory network with attention, capturing sequential dependencies in text [16].
- $BERT_t$: BERT-base-uncased fine-tuned on FakeNewsNet without the lexicon fusion, serving as the upper-bound ablation baseline to quantify the added value of the emotion module.

4.4. Implementation Details

All experiments are implemented in Python 3.10 using PyTorch 2.1 and the Hugging Face Transformers library. BERT fine-tuning is performed on an NVIDIA RTX 3080 GPU (10 GB VRAM). TextBlob version 0.17.1 is used for subjectivity and polarity extraction. The NRC Emotion Lexicon (English, version 0.92) is queried through the NRCLex Python wrapper. All experiments use random seed 42 for reproducibility. Training time for BERT fine-tuning is approximately 22 minutes per run.

5. Results and Discussion

5.1. Classification Performance

Table 2 reports classification performance on the FakeNewsNet test set. The proposed model achieves the best result among all evaluated systems, with an F_1-score of 0.77 and an AUC of 0.84. It outperforms the standalone $BERT_t$ baseline ($F_1 = 0.74$, AUC = 0.81) by 3 F_1 points, demonstrating that the emotion-aware fusion adds measurable value beyond the transformer classifier alone. The BiLSTM model achieves moderate performance ($F_1 = 0.69$), confirming that sequential architectures without contextual pre-training fall short of transformer-based approaches on this task.

Table 2.
Classification performance on the FakeNewsNet test set (n = 558).

Model	Precision	Recall	F_1	AUC
TF-IDF + SVM Rashkin, et al. [13]	0.63	0.61	0.62	0.66
CNN Kim [15]	0.67	0.65	0.66	0.70
BiLSTM Hochreiter and Schmidhuber [16]	0.70	0.68	0.69	0.73
$BERT_t$ Devlin, et al. [10]	0.75	0.73	0.74	0.81
Proposed Model	0.78	0.76	0.77	0.84

For cross-dataset generalisation, the proposed model achieves an F_1-score of 0.71 on the LIAR dataset without any domain-specific retraining, compared to 0.68 for $BERT_t$ alone. The consistent improvement across both datasets suggests that the emotion signal is domain-agnostic, capturing a fundamental property of deceptive content rather than dataset-specific artefacts.

5.2. Ablation Study

Table 3 presents the results of an ablation study in which each component of the proposed model is removed in turn. The NRC emotion module (Stage 2) contributes the largest performance gain: removing it causes a drop of 4 F_1 points (from 0.77 to 0.73). This confirms the central hypothesis that emotional language is a strong and robust indicator of misinformation. The subjectivity/polarity module (Stage 1) contributes a smaller but consistent gain (+2 F_1 points), while the BERT component (Stage 3) is the single most powerful individual predictor ($F_1 = 0.74$ standalone) but benefits from the complementary signals provided by the lexicon modules.

Table 3.
Ablation study on FakeNewsNet test set.

Configuration	F_1	AUC	ΔF_1
Full model (S_1 fusion)	0.77	0.84	—
w/o Stage 1 (no TextBlob)	0.75	0.82	−0.02
w/o Stage 2 (no NRC emotions)	0.73	0.79	−0.04
w/o Stage 3 (no BERT)	0.71	0.76	−0.06
$BERT_t$ only (standalone)	0.74	0.81	−0.03

5.3. Emotional Marker Analysis

Table 4 reports the mean NRC emotion scores for fake and real articles in the test set. Fake articles exhibit significantly higher levels of anger ($\mu = 0.18$ vs. 0.07), fear ($\mu = 0.15$ vs. 0.06), and disgust ($\mu = 0.12$ vs. 0.04), consistent with prior findings in the literature [5, 22]. Conversely, real articles score higher on trust ($\mu = 0.21$ vs. 0.09) and anticipation ($\mu = 0.14$ vs. 0.08). This emotional asymmetry justifies the design choice to assign higher weight to the NRC component in the fusion formula.

Table 4.
Mean NRC emotion scores by article category (test set, n = 558).

Emotion	Fake News (μ)	Real News (μ)	Difference
Anger	0.18	0.07	+0.11*
Fear	0.15	0.06	+0.09*
Disgust	0.12	0.04	+0.08*
Sadness	0.10	0.08	+0.02
Trust	0.09	0.21	−0.12*
Anticipation	0.08	0.14	−0.06*
Joy	0.07	0.11	−0.04
Surprise	0.09	0.09	0.00

Note: * Statistically significant difference (Mann-Whitney U test, $p < 0.01$).

5.4. Discussion

The experimental results confirm four main conclusions. First, emotional language is a robust and domain-agnostic signal for fake news detection. The NRC emotion module contributes the largest marginal gain in the ablation study, and

the emotional profiles in Table 4 match theoretical predictions from Plutchik's psychoevolutionary model [24]. This robustness across the political (LIAR) and mixed-media (FakeNewsNet) domains suggests that the emotionally manipulative rhetorical strategy of misinformation transcends topical domains. Second, lexicon-based features and contextual transformer representations capture complementary aspects of the text. The NRC module identifies explicit emotional markers, while BERT captures implicit semantic patterns not expressed through individual charged words — such as irony, understatement, and manipulative framing. This complementarity explains why the full fusion systematically outperforms any individual component, including standalone BERT fine-tuned. Third, the lightweight design makes the model practical for resource-constrained deployment. The TextBlob and NRC components require no GPU and add negligible latency at inference, while BERT can be served from a modest GPU or CPU with quantisation, matching the computational infrastructure typically available in Sub-Saharan African deployment contexts. Fourth, the interpretability of the credibility score S_1 and the underlying emotion breakdown supports media literacy development in addition to automated flagging, an important factor for user adoption in communities with limited trust in automated systems.

Two limitations deserve explicit mention. The NRC Emotion Lexicon was constructed primarily from English text and may not capture emotion expressions in French, Fulfulde, or Camfranglais, the three main languages of the Cameroonian information ecosystem. Extending the approach to these languages requires either translation-based lexicon adaptation or the development of new multilingual emotion resources. Additionally, the model handles text only and cannot process the images, videos, or audio that increasingly circulate in WhatsApp misinformation campaigns. Future work will address both limitations through multilingual lexicon extension and multimodal feature integration.

6. Conclusion

This paper presented an emotion-driven credibility scoring model for fake news detection combining three complementary components: subjectivity and polarity features from TextBlob, eight-dimensional emotion intensity profiles from the NRC Emotion Lexicon, and a contextual classification probability from a fine-tuned BERT-base-uncased model. The three signals are fused through a weighted formula calibrated by grid search on validation data, yielding a credibility score $S_1 \in [0,1]$ that is both principled and interpretable.

Experiments on FakeNewsNet demonstrate $F_1 = 0.77$ and AUC = 0.84, outperforming all evaluated baselines including standalone fine-tuned BERT ($F_1 = 0.74$). The ablation study confirms that the NRC emotion module provides the largest marginal gain, with anger, fear, and disgust identified as the most discriminative emotional markers. Cross-domain validation on LIAR ($F_1 = 0.71$ without retraining) confirms the robustness of the emotion signal across topical domains. Statistical analysis of NRC emotion profiles reveals that fake articles exhibit significantly higher anger, fear, and disgust levels, while real articles score higher on trust and anticipation, consistent with Plutchik's psychoevolutionary model [24] and prior literature [5, 22]. The model is computationally affordable and has been integrated into the EyeAndEars platform for deployment in the Cameroonian digital context. Its interpretability — the credibility score and emotion breakdown are human-readable — makes it suitable for supporting media literacy programmes in addition to automated flagging.

Three future research directions are prioritised. First, extending the NRC lexicon to French, Fulfulde, and Camfranglais through cross-lingual transfer or direct crowdsourced annotation, to enable native deployment in the multilingual Cameroonian information ecosystem. Second, incorporating source-level credibility signals through automated web crawling to complement the text-level analysis, addressing factual verification beyond rhetorical style. Third, collecting and annotating a locally sourced Cameroonian fake news dataset to train and evaluate detection systems specifically tailored to the regional linguistic and cultural context.

References

[1] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146-1151, 2018, https://doi.org/10.1126/science.aap9559.

[2] M. S. Islam *et al.*, "COVID-19–related infodemic and its impact on public health: A global social media analysis," *The American Journal of Tropical Medicine and Hygiene*, vol. 103, no. 4, p. 1621, 2020, https://doi.org/10.4269/ajtmh.20-0812.

[3] C. Wardle and H. Derakhshan, *Information disorder: Toward an interdisciplinary framework for research and policymaking*. Strasbourg: Council of Europe, 2017.

[4] O. D. Apuke and B. Omar, "Fake news and COVID-19: Modelling the predictors of fake news sharing among social media users," *Telematics and Informatics*, vol. 56, p. 101475, 2021, https://doi.org/10.1016/j.tele.2020.101475.

[5] B. Horne and S. Adali, "This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, no. 1, pp. 759-766, 2017.

[6] X. Zhang, J. Cao, X. Li, Q. Sheng, L. Zhong, and K. Shu, "Mining dual emotion for fake news detection," in *Proceedings of the Web Conference*, 2021.

[7] V. Perez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic detection of fake news," in *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, 2018.

[8] S. Loria, "TextBlob: Simplified text processing, version 0.17.1," Retrieved: https://textblob.readthedocs.io/. 2021.

[9] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word–emotion association lexicon," *Computational Intelligence*, vol. 29, no. 3, pp. 436-465, 2013, https://doi.org/10.1111/j.1467-8640.2012.00460.x.

[10] J. Devlin, M. 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 American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1 (long and short papers)*, 2019.

[11] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," *Big Data*, vol. 8, no. 3, pp. 171-188, 2020, https://doi.org/10.1089/big.2020.0062.

[12] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th International Conference on World Wide Web*, 2011.

[13] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of varying shades: Analyzing language in fake news and political fact-checking," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017.

[14] W. Y. Wang, "Liar, liar pants on fire": A new benchmark dataset for fake news detection," in *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017.

[15] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

[17] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "dEFEND: Explainable fake news detection," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '19)*, 2019.

[18] J. Ma, W. Gao, and K. F. Wong, "Rumor detection on twitter with tree-structured recursive neural networks," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018.

[19] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019, https://doi.org/10.48550/arXiv.1907.11692.

[20] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019, https://doi.org/10.48550/arXiv.1910.01108.

[21] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," *Advances in Neural Information Processing Systems*, vol. 32, pp. 5753–5763, 2019.

[22] A. Giachanou, P. Rosso, and F. Crestani, "Leveraging emotional signals for credibility detection," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 877–880). Association for Computing Machinery*, 2019.

[23] S. Singhania, N. Fernandez, and S. Rao, *3han: A deep neural network for fake news detection in international conference on neural information processing*. Cham: Springer International Publishing, 2017.

[24] R. Plutchik, *A general psychoevolutionary theory of emotion," in theories of emotion, R. Plutchik and H. Kellerman, Eds*. New York: Academic Press, 1980.

[25] F. Tchakounté, A. Faissal, M. Atemkeng, and A. Ntyam, "A reliable weighting scheme for the aggregation of crowd intelligence to detect fake news," *Information*, vol. 11, no. 6, p. 319, 2020.

[26] M. Honnibal and I. Montani, "SpaCy 2: Natural language understanding with bloom embeddings," Convolutional Neural Networks and Incremental Parsing, 2017.