

Context-Aware And Explainable Hate Speech Detection In Iraqi Social Media: A Framework For Parliamentary And Policy Analysis

Ali N. Al-Saadi

Department of Parliamentary Research and Studies, Iraqi Council of Representatives, Iraq

Received: 29 Aug 2025 | Received Revised Version: 13 Sep 2026 | Accepted: 28 Sep 2026 | Published: 16 Oct 2026

Volume 08 Issue 10 2026 | DOI:- 10.37547/tajssei/Volume08Issue10-02

Abstract

Overview: Concerns about hate speech and its implications for social cohesion, public discourse, and the safety of targeted communities have been intensified by the rapid expansion of social media, particularly in Iraq owing to its social, cultural, and linguistic diversity, as well as its experience of severe sectarian conflict and communal violence during 2006–2007. Hence, it is important to examine social cohesion and harmful intergroup discourse in contemporary digital environments without assuming a direct causal relationship between past conflict and present-day online hate speech.

Gap: Machine learning and Transformer-based hate-speech detection frameworks have significantly advanced in recent times; however, the reliable identification of hate speech in Iraqi social-media discourse remains challenging since online communication in the country often involves dialectal variation, informal and non-standard spelling, code-mixing, sarcasm, implicit and coded expressions, and deliberate linguistic obfuscation. These context-dependent and dialect-sensitive characteristics may not be adequately captured by generic language representations, and classification performance alone may fail to provide enough insight into model explainability, generalization across different forms of Iraqi online discourse, and robustness to linguistic manipulation. Hence, there is a need for a framework that can combine contextual understanding, dialect sensitivity, explainability, and robustness for the Iraqi context.

Contributions: A context-aware and explainable hate-speech detection framework has been proposed in this study for deployment across Iraqi social media; the methodology of the proposed framework dynamically combines contextual and dialect-specific information for hate-speech classification by integrating Iraqi-aware text preprocessing, contextual Transformer representations, dialect-sensitive feature learning, and an adaptive fusion mechanism. The system also incorporates Explainable Artificial Intelligence (XAI) for identifying the linguistic elements contributing to model predictions; robustness analysis evaluates model behavior under realistic linguistic variations and obfuscation strategies. In addition to the computational contribution, an interdisciplinary AI–policy framework is also incorporated, wherein anonymized and aggregated hate-speech patterns can provide empirical evidence for parliamentary and public-policy analysis. The legislative and oversight role of the Iraqi Council of Representatives in addressing online hate speech was also examined in consideration of freedom of expression, privacy, explainability, responsible digital governance, and human oversight.

Conclusion: An integrated approach that connects Iraqi-contextualized hate-speech detection with explainable and robust artificial intelligence and parliamentary policy analysis has been proposed in this study, wherein AI is positioned as an empirical research and decision-support instrument rather than a mechanism for automated censorship, punishment, or legal determination. A combination of linguistic adaptation, contextual modeling, explainability, robustness assessment, and aggregated policy-oriented evidence in the proposed methodology provides an experimentally testable and socially responsible foundation for analyzing hate speech in Iraqi social media and supporting evidence-informed parliamentary and public-policy analysis.

Keywords: Hate speech detection; Iraqi Arabic; Social media; Natural language processing; Transformer models; Explainable artificial intelligence; Context-aware learning; Iraqi Council of Representatives; Digital governance; Iraq.

© 2026 Ali N. Al-Saadi. This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). The authors retain copyright and allow others to share, adapt, or redistribute the work with proper attribution.

Cite This Article: Ali N. Al-Saadi. (2026). Context-Aware And Explainable Hate Speech Detection In Iraqi Social Media: A Framework For Parliamentary And Policy Analysis. The American Journal of Social Science and Education Innovations, 8(10), 10–29. <https://doi.org/10.37547/tajssei/Volume08Issue10-02>

1. Introduction

Information exchange between people and communities has been greatly transformed by the rapid expansion of social media, enabling people to express opinions and participate in public discourse at unprecedented speed and scale. This same digital environment, however, has facilitated the dissemination and amplification of hostile, discriminatory, and hateful expressions, such as hate speech. Scholars have defined hate speech as forms of communication that employ discriminatory language toward individuals or groups based on identity-related characteristics, such as religion, ethnicity, nationality, race, or gender [1]. In many cases, such discourse reinforces discrimination and marginalization, thereby affecting social cohesion and people's safety [1], [2]. Hence, it has become important to detect and understand online hate speech, making it an important research problem at the intersection of Natural Language Processing (NLP), Machine Learning (ML), Deep Learning (DL) [3], and social-media analytics. However, it is problematic to automatically detect hate speech since natural language is highly contextual, and similar meanings can be expressed through substantially different linguistic constructions. Researchers have developed neural architectures and contextual Transformer-based language models, moving beyond conventional lexical and machine-learning techniques. For instance, Transformer models such as BERT can generate bidirectional contextual representations by interpreting linguistic elements in relation to their surrounding context [4]. Contextual representation, on its own, cannot fully resolve issues associated with dialectal variation, coded expressions, non-standard language, implicit meaning, or the distinction between hateful and merely offensive discourse.

This study considers Iraq as an important case study for hate-speech research because contemporary digital communication occurs within a socially, culturally, and linguistically diverse environment and a historically

significant context of intergroup relations. Again, Iraq experienced severe sectarian conflict and communal violence between 2006 and 2007, where the United Nations documented a substantial deterioration in the security situation, increasing sectarian violence, civilian casualties, displacement, and intercommunal tensions in 2006 [5]. Further reports by the United Nations in 2006 documented prospects of an all-out civil war becoming substantially more serious following intensification in sectarian violence and insecurity [6]. Therefore, this historical period provides important societal context for investigating social cohesion and intergroup discourse in Iraq. This should, however, not be interpreted as evidence that the conflict of 2006-2007 is a direct cause of contemporary online hate speech in Iraq.

Institutional attention has been given to the relevance of hate speech in the contemporary digital context in Iraq. In 2023, the UNAMI Human Rights Office held a dialogue on the role of the media and social media in the fight against hate speech in Iraq [5]. Discussions included stakeholders from governmental institutions, media coverage, civil society, religious and ethnic organizations, journalists, activists, and content on social media. Hate speech, minority perspectives, social cohesion, relations with media and social-media platforms, and countering hate speech and protecting freedom of expression were discussed [5]. These concerns are an indication that it is important to carefully examine harmful online discourse without assuming that it cannot be found in some communities, or is ubiquitous.

In computing terms, the communication of Iraqi people via social media is an interesting linguistic problem. Texts of online Arabic may include Modern Standard Arabic, Iraqi Arabic varieties, code-mixed expressions, and sentences with Arabic–English, regional linguistic variation, Arabic–English code-mixing, Arabic–English utilizing Arabic script (Arabizi), hashtags and emojis, slang and sarcasm, coded expressions, and hostile expressions (explicit or

implicit). This language deviation can be challenging for automated classification to contend with, as seemingly identical phrases might mean something completely different through dialect and context, and hateful meaning may not be explicitly represented as hate speech. Hence, it is necessary to determine the occurrence and prevalence of these linguistic phenomena empirically, not a priori.

While Transformer-based approaches have greatly improved text classification and hate-speech detection systems, the detection of hate speech in the context of Iraqi social media is still difficult to achieve. A single, generic multilingual representation might not be the most effective method for conveying characteristics specific to the Iraqi dialect and culture, especially when the meaning intended to be conveyed is obscured, used in coded terms, in the informal language of Arabic, or through intentional linguistic adaptation. Furthermore, the accuracy of classification does not provide strong evidence with respect to robustness against linguistic variations, the interpretability of its predictions, and its ability to generalize to discourses that it has not seen before. Another fundamental problem is whether offensive speech constitutes group-targeted hate speech or not, as offensive speech does not necessarily imply group-targeted hate speech. More attention should, therefore, be paid to an Iraq-specific computational framework, which integrates contextual and dialect-specific representation of language with explainability, robustness, and generalization.

This study explores the need for a context-aware, explainable Hate Speech Detection ML model for the Iraqi social media context. The proposed approach aims to study the effect of embedding contextual and dialect-sensitive linguistic information to produce an effective representation of complex Iraqi online discourse without confusing hateful, offensive, and non-hateful expressions under the adopted annotation framework.

The research will also seek to move beyond simply evaluating the model's predictive performance to include evaluating models with post-predictive tasks such as explainability, robustness, error analysis, and generalization assessments. These dimensions allow for the study of both the actual ability of the model to detect hate speech and which linguistic features impact its predictions, how its performance changes with real-world linguistic variations and obfuscation, and the ability to (possibly) generalize beyond the set of data used for training the model. This study significantly contributes to both the computational

and interdisciplinary aspects of addressing hate speech, including the intersection of artificial intelligence and research in politics and public policy within parliament. Furthermore, computational responses to hate-speech analysis need to be considered in the context of freedom of expression, privacy, false-positive and false-negative errors, model bias, and explainability.

The study focuses on the possibility of analyzing Internet-connected legislative and oversight mechanisms available to the Council of Representatives of Iraq, taking into account the anonymized and aggregated evidence generated by AI analysis of patterns of hate speech online. AI is not seen as an automatic censor, as a tool to automatically judge the criminality of citizens, or as a tool for automatically targeting citizens for punishments, or as an alternative to legislation and judicial decisions; rather, AI is considered to be a research and decision-making tool and not to automatically censor citizens, determine criminal responsibility, select citizens for punishments, or act as a substitute for legislative and judicial decisions.

This study explores four major research contributions;

The methodological contribution focuses on the creation of a context-aware and dialect-sensitive approach to the detection of hate speech in the context of Iraqi social media discourse. The second part of the linguistic aspect is the systematic research of hate-speech recognition in Iraqi online communication.

Systematic research of hate-speech recognition in Iranian online communication; the linguistic aspect of the project requires systematic research of hate-speech recognition in Iranian online communication. This study takes a less “universal” perspective on general Arabic representations, and considers the linguistic forms that will be observed empirically (dialectal, contextual, informal, coded, and potentially obfuscated).

The trustworthy-AI contribution goes beyond the typical category evaluation metrics (explaining, robustness testing, ablation studies, error analyses, and generalization experiments). The analyses aim to characterize the contribution of the elements of the computation to the behavior of the model, the linguistic evidence it maps onto to make predictions, the way in which the model reacts to realistic linguistic perturbations, and whether its performance can be generalized to other contexts.

The parliamentary and public-policy contribution discusses simulation evidence secured using AI-driven hate-speech analysis that could be helpful in studying the Legislative and Oversight Function of the Council of Representatives of the Republic of Iraq. This dimension addresses computational

hate speech analysis and related matters of social cohesion, constitutional rights, freedom of expression, privacy, digital accountability, and human oversight without taking a political or legislative stance.

Hate speech detection on Iraqi social media is a multi-faceted research challenge that is not only linguistically complex but also demands computational reliability as well as social and public-policy concerns. The importance of social cohesion and intergroup discourse, as well as the continuity and ongoing problems of harmful communication in Iraq's digital environment, have been illustrated by the severe sectarian conflict and communal violence that occurred in 2006–2007, and by the recent institutional efforts to tackle the issues of hate speech and social media in Iraq [7]. The historical context motivated the present study, but is never given as evidence that there is a causal link between the past conflict and current hate speech in online communication. Therefore, this study establishes an integrated research direction combining Iraqi-contextualized hate-speech detection, dialect-sensitive language representation, explainable and robust AI, and parliamentary and public-policy analysis. Section 2 details the computational architecture, Iraqi-aware preprocessing procedures, mathematical formulation, classification mechanism, explainability techniques, robustness mechanisms, and experimental protocol. Section 3 covers the results and discussions, while the conclusion, limitations, and future directions are provided in Section 4.

2. PROPOSED METHODOLOGY

In this work, the aim is to design and develop a framework for hate speech detection in Iraqi social-media discourse; the proposed framework is also intended to address contextual semantics, dialectal variation, informal linguistic forms, and intentionally obfuscated expressions associated with Iraqi social media discourse. The proposed framework, unlike a conventional text-classification pipeline that relies primarily on a single pretrained contextual representation, combines a contextual Transformer representation with an Iraqi dialect-sensitive representation through an adaptive fusion mechanism. XAI and robustness assessment are subsequently incorporated to provide interpretable predictions and evaluate model behavior under realistic linguistic variations. Furthermore, the proposed methodology maintains a clear distinction among hate speech, offensive language, and non-hateful discourse whenever this three-class formulation is supported by the annotation framework.

3. Waterfall Methodology

The Waterfall methodology forms the basis for the development of the proposed hate-speech detection framework in this study; the process consists of requirement analysis, system design, implementation, experimental testing, evaluation, and refinement. The first stage involves identification of the computational and linguistic requirements of Iraqi hate-speech detection, such as the representation of Iraqi Arabic varieties, contextual semantics, informal spelling, code-mixing, Arabizi, implicit expressions, coded language, and other linguistic phenomena observed empirically in the collected corpus. The design stage involved the specification of the computational components of the proposed framework, including Iraqi-aware preprocessing, contextual Transformer encoding, dialect-sensitive representation, adaptive feature fusion, hate-speech classification, explainability analysis, and robustness evaluation. During the implementation stage, these components are translated into an executable deep-learning framework; the testing phase covers model evaluation using held-out data and predefined evaluation metrics. The last phase is model refinement using training and validation information; the test set remains isolated from model selection and hyperparameter tuning to preserve unbiased final evaluation.

4. Machine Learning Process

Step 1. Data Acquisition

In this stage, a suitable corpus of Iraqi social-media discourse is obtained; first, an existing Iraqi hate-speech corpus is checked for satisfying the linguistic, annotation, accessibility, and ethical requirements of the study; if satisfactory, its provenance, class labels, linguistic coverage, annotation procedure, and accessibility are documented; otherwise, an Iraqi social-media corpus will be constructed using an ethically approved collection strategy. Such data must represent Iraqi online communication and may include Modern Standard Arabic, Iraqi Arabic varieties, informal Iraqi Arabic, code-mixed Arabic-English text, Arabizi, hashtags, emojis, slang, and other linguistic phenomena where empirically observed. The presence of these features will be determined from the corpus rather than assumed beforehand. The research dataset will be anonymized according to the adopted ethical protocol, and user-identifying information will not constitute a predictive feature. When supported by the annotation framework, the label space used when the annotation protocol supports three classes is given as follows:

$$Y = \{\text{Hate, Offensive, Neither}\} \quad (1)$$

Where Y = a set of possible target labels;

'Hate' = group-targeted hateful expression under the study's operational definition;

'Offensive' = abusive or insulting language that does not satisfy the hate-speech definition;

'Neither' = content that belongs to neither category.

In the context of this study, hate speech is distinguished from general offensive language because offensive language does not necessarily constitute group-targeted hate speech. Corpus annotation will involve multiple independent annotators; disagreements will be resolved through a predefined adjudication procedure, and inter-annotator agreement will be reported. Hate targets, where supported by the annotation scheme, may additionally be categorized into predefined research categories without inferring sensitive individual users' features.

Step 2. Data Cleaning and Iraqi-Aware Preprocessing

Social-media text must first undergo a preprocessing step before representation learning; however, aggressive normalization may remove linguistic signals that contribute to hate-speech interpretation. Hence, Iraqi-aware preprocessing is considered in the proposed methodology rather than indiscriminately deleting non-standard textual elements. Iraqi-aware preprocessing is captured as follows:

$$X = P_Iraq(X_raw) \quad (2)$$

Where

X_raw = the original social-media post before normalization,

$P_Iraq(\cdot)$ = the preprocessing transformation designed for the linguistic properties observed in Iraqi online communication. The output, X , is the cleaned but semantically preserved input fed to the model.

The transformation may normalize selected Arabic orthographic variants and remove non-informative artifacts, while retaining potentially meaningful signals such as dialectal spellings, hashtags, emojis, punctuation, repetition, Arabizi, and code-mixed expressions when justified by the data. The preprocessing process evaluates normalization of Arabic Alef variants, Ya/Alif Maqsura forms, Tatweel, diacritics, excessive whitespace, URLs, and user mentions. Repeated characters, when identified, are cautiously normalized since character repetition may

convey emphasis. Features such as hashtags, punctuation, emojis, Iraqi colloquial spellings, Arabic-Latin mixing, Arabizi, and English code-mixed expressions are preserved or transformed into informative representations when carrying semantic information. Information leakage is minimized by identifying duplicate and near-duplicate records before splitting the corpus; user-level and temporal leakage are also controlled where metadata permits. Each text instance, after preprocessing, is given as:

$$X = (x_1, x_2, \dots, x_n) \quad (3)$$

Where

x_i = the token at position i after tokenization, etc.

Transformers use relationships among tokens to construct contextual meaning; hence, token order preservation is important. The effective sequence length after preprocessing, tokenization, and any model-specific truncation or padding is represented by the value n .

Step 3. Choosing the Model

In language models based on transformers, the values of tokens are conditioned on their linguistic context. BERT added a new dimension of deep bidirectional Transformer representations which are customizable for downstream NLP tasks [8], [9]. The former models, designed for exclusive use of the Arabic language, are Arabic-specific models like AraBERT [10], and the latter are multilingual models like XLM-R [11] designed to handle both multilingual and code-mixed environments. The final Transformer backbone will be selected experimentally rather than assumed to be superior in advance.

The proposed model in this study is called Context-Aware Iraqi Dialect Explainable Hate-Speech Network (CID-XHS), which consists of two complementary representation branches: a Hate-Speech Representation branch, which is a Context-Aware Transformer, and a Representation branch that is Iraqi Dialect-Sensitive.

1) Contextual Representation

The main contextual-encoding operation is represented as:

$$H_c = E_{\theta}(X) \quad (4)$$

Where

E_{θ} = the selected pretrained Transformer encoder;

θ = its trainable parameters.

When the encoder is fed with the token sequence X , the output H_c contains contextualized token representations;

each representation depends on the surrounding tokens, allowing the same lexical item to receive different representations in different contexts.

Equation (5) expands H_c into its n contextual token states as follows:

$$H_c = (h_1, h_2, \dots, h_n) \quad (5)$$

Where

Each h_i = a d -dimensional hidden representation associated with token x_i after Transformer encoding.

These vectors collectively retain token-level contextual information and form the input to the subsequent attention-pooling mechanism.

An unnormalized attention relevance score e_i for contextual state h_i is computed as follows:

$$e_i = v_a^T \tanh(W_a h_i + b_a) \quad (6)$$

Where

W_a and b_a are learnable projection parameters,

$\tanh$ introduces non-linearity,

v_a is a learnable context vector that converts the transformed hidden state into a scalar score (a larger e_i implies that, based on the learned parameters, token i is assigned greater relevance for constructing the message-level contextual representation).

The raw attention scores are converted into normalized attention weights α_i using the softmax function as in Equation (7):

$$\alpha_i = \exp(e_i) / \sum_{j=1}^n \exp(e_j) \quad (7)$$

The exponential makes all weights positive, while division by the sum across the n tokens ensures that the weights sum to one. Hence, α_i can be interpreted as the relative contribution assigned by this pooling mechanism to token i when forming the contextual summary.

The formation of the message-level contextual vector h_c as a weighted sum of the token states is represented in Equation (8):

$$h_c = \sum_{i=1}^n \alpha_i h_i \quad (8)$$

Where

Each hidden state h_i is multiplied by its normalized attention weight α_i ; tokens that receive higher learned

weights therefore contribute more strongly to h_c . With this, the variable-length contextual sequence is compressed into a fixed-dimensional representation suitable for fusion and classification.

2) Iraqi Dialect-Sensitive Representation

The complementary Iraqi dialect-sensitive representation is given in Equation (9):

$$h_d = D_\phi(X) \quad (9)$$

Where

D_ϕ = a learnable representation function parameterized by ϕ ,

h_d = its output for the same input X (its purpose is to encode dialect-sensitive information observed in the corpus, such as Iraqi colloquial forms, regional variants, non-standard spellings, Arabizi, and code-mixing). The usefulness of this branch is tested through ablation experiments, not assumed.

3) Adaptive Fusion

The contextual representation h_c and dialect-sensitive representation h_d are concatenated into a joint vector u as in Equation (10):

$$u = [h_c ; h_d] \quad (10)$$

Where

The semicolon represents vector concatenation rather than arithmetic addition; this operation exposes both information sources to the gating network, enabling the model to learn the pattern of changes in their relative usefulness across inputs.

The adaptive gate g is computed as given in Equation (11):

$$g = \sigma(W_g u + b_g) \quad (11)$$

Where

W_g and b_g = trainable parameters,

σ = the sigmoid function that maps the gate values to the interval (0,1); the gate is conditioned on the concatenated vector u ; therefore, it can vary from one social-media post to another, where values closer to one favor the contextual branch, while values closer to zero increase the relative contribution of the dialect-sensitive branch, dimension by dimension when g is vector-valued.

The fused representation z is produced as given in Equation (12);

$$z = g \odot h_c + (1 - g) \odot W_d h_d \quad (12)$$

Where

$\odot$ = element-wise multiplication.

$g \odot h_c$, weights the contextual features;

$(1-g) \odot W_d h_d$, weights the dialect-sensitive features after W_d projects h_d into a compatible feature space. Considering the complementary nature of the two weights, the fusion mechanism learns an input-dependent balance between general contextual semantics and Iraqi/dialect-sensitive evidence.

4) Hate-Speech Classification

The fused representation z is converted into class probabilities using Equation (13);

$$P(y|X) = \text{Softmax}(W_c z + b_c) \quad (13)$$

Where

W_c and b_c = parameters of the final classification layer whose linear transformation produces one logit for each class; Softmax normalizes these logits into $P(y|X)$, a probability distribution whose values sum to one. The components correspond to Hate, Offensive, and Neither under the three-class formulation.

The final predicted label $\hat{y}$ is given by Equation (14):

$$\hat{y} = \arg \max_y P(y|X) \quad (14)$$

Where the $\arg \max$ operator selects the class y having the largest posterior probability in $P(y|X)$; this is the model's classification decision, which on its own does not constitute a legal, social, or causal determination about the author of the input.

The model is optimized for multiclass classification using categorical cross-entropy, given as follows:

$$L_{CE} = -(1/N) \sum_{i=1}^N \sum_{k=1}^K y_{ik} \log(p_{ik}) \quad (15)$$

Where

N = the number of training instances,

K = the number of classes,

y_{ik} = the ground-truth indicator,

p_{ik} = the predicted probability. If class imbalance is empirically observed, class-weighted loss will be evaluated rather than assuming beforehand that balancing is required.

The equation defines categorical cross-entropy loss for N training examples and K classes, where y_{ik} equals one when example i belongs to class k and zero otherwise; p_{ik} is the model-assigned probability to that class. The negative logarithm penalizes low probability assigned to the correct class. Therefore, minimizing L_{CE} encourages an increase in the model's predicted probability of the ground-truth class. To ensure minority classes are not underrepresented in the optimization objective, a weighted form can be evaluated if the empirical corpus is imbalanced.

Step 4. Model Training

Stratification will be used where appropriate to separate the dataset into training, validation, and test sets; duplicate, user-level, and temporal leakage will be controlled wherever permitted by the available metadata. Model parameters will be optimized using the training set, while the validation set will support model selection and hyperparameter tuning; final model evaluation will rely solely on the test set.

$$\Theta_{t+1} = \Theta_t - \eta \nabla_{\Theta} L \quad (16)$$

The generic gradient-based parameter update during training is expressed in Equation (16), where Θ denotes the trainable parameters, η is the learning rate, and L is the optimization objective; Θ_t is the complete set of trainable parameters at iteration t , $\nabla_{\Theta} L$ is the gradient of the loss with respect to those parameters, and η is the learning rate; to reduce the training objective, the update moves the parameters in the direction opposite to the gradient. During implementation, an optimizer such as AdamW may realize a more sophisticated version of this principle; however, the exact optimizer and settings must be reported from the actual experiments.

The final implementation will report the Transformer checkpoint, tokenizer, maximum sequence length, optimizer, learning rate, batch size, epochs, early-stopping criterion, weight decay, class weighting, random seeds, hardware, and software versions; these values will be selected during implementation and validation.

Step 5. Model Evaluation

Evaluation of the proposed framework will be done using Accuracy, Precision, Recall, F1-score [12], [13], [14], [15], Macro-F1, Weighted-F1 [16], [17], and Matthews Correlation Coefficient (MCC) [18]; Macro-F1 will be

considered when there is an imbalanced empirical class distribution.

Precision is a measure of the proportion of instances predicted as that class that are actually members of the class. In hate-speech detection, low precision corresponds to many non-hateful instances being incorrectly labeled as hate, making this metric particularly relevant to false-positive analysis. Precision is given as:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (17)$$

Where;

TP = true positives

FP = false positives for the class under evaluation.

Recall is a measure of the proportion of actual positive instances that the classifier successfully identifies; for the Hate class, low recall means that many truly hateful examples are missed. Precision is given as:

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (18)$$

Where FN denotes false negatives.

Precision and Recall capture different error types and should be considered together.

F1-score is the harmonic mean of Precision and Recall; it is given as:

$$\text{F1} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (19)$$

The harmonic mean decreases when either component is low; hence, F1 rewards models that maintain a balance between false-positive control and the ability to retrieve positive instances. For multiclass evaluation, F1 can be calculated per class and then aggregated using macro or weighted averaging.

Accuracy refers to the proportion of correct predictions among all evaluated instances in the binary notation shown. It is given as:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (20)$$

Accuracy, for multiclass experiments, is equivalently the number of correctly classified instances divided by the total number of instances. Therefore, accuracy is intuitive but can be misleading under class imbalance, which is why the methodology also reports Macro-F1 and other class-sensitive measures.

Per-class metrics and a confusion matrix will be reported for the multiclass task; ROC-AUC, PR-AUC, false-positive rate, and false-negative rate will also be considered where appropriate. Comparison will be done with TF-IDF + Logistic Regression, TF-IDF + SVM, CNN, BiLSTM, BiLSTM + Attention, BERT, mBERT, an Arabic Transformer, XLM-R, and a standard BERT-based classifier.

Step 6. Parameter Tuning and Ablation Analysis

The training and validation sets will be employed for hyperparameter optimization; the considered parameters for tuning include learning rate, batch size, maximum sequence length, dropout probability, training epochs, weight decay, attention dimensions, and adaptive-fusion dimensions.

M_Full, the complete proposed CID-XHS configuration, is denoted in Equation (21);

$$\text{M_Full} \quad (21)$$

It contains the selected Transformer representation, dialect-sensitive branch, adaptive fusion, classification component, and the robustness-related training components that are actually implemented. M_Full serves as the reference configuration for comparing the ablated variants.

The principal ablation variants are defined in Equation (22);

$$\text{M_Dialect}, \text{M_Fusion}, \text{M_Robust}, \text{M_Transformer} \quad (22)$$

Where

M_Dialect removes the dialect-sensitive branch;

M_Fusion removes or replaces adaptive fusion;

M_Robust excludes robustness-oriented training if such training is implemented;

M_Transformer reduces the system to the base Transformer classifier.

These variants are compared with M_Full to identify the empirical contribution of individual architectural choices instead of attributing any observed improvement to the full system without evidence. The full model will be compared with variants that remove the dialect-sensitive representation, adaptive fusion, robustness-oriented training where implemented, and the added components beyond the base Transformer. This makes the contribution of each proposed component experimentally testable.

Step 7. Explainability, Robustness, and Final Prediction

The explainability aspect of the evaluation stage investigates input elements for contribution to model predictions; Integrated Gradients (IG) is adopted as a principal attribution method as it provides feature-level attribution for differentiable models [19].

Integrated Gradients for input feature I is given in Equation (23);

$$IG_i(x) = (x_i - x'_i) \int_0^1 [\partial F(x' + \alpha(x - x')) / \partial x_i] d\alpha \quad (23)$$

Where

x = the actual input representation,

x' = a baseline representation,

F = the model output being explained,

α continuously interpolates between the baseline and the actual input.

The derivative measures the level of sensitivity of F to feature x_i along that path, and the integral accumulates this sensitivity from $\alpha=0$ to $\alpha=1$; multiplication by $(x_i - x'_i)$ scales the accumulated gradient by the feature's displacement from the baseline. The integral is approximated numerically in practice, and the resulting IG_i value is an attribution of model behavior, not evidence that the corresponding token causally produces hate speech in society.

Linguistic elements that influence a model prediction are identified from the resulting token-level attribution scores; these attributions are interpreted as explanations of model behavior rather than evidence of linguistic or social causality.

Controlled perturbations that reflect realistic social-media variation, including misspellings, character substitutions, altered spacing, repeated characters, Arabizi, dialect substitutions, obfuscated terms, emoji substitution, and paraphrasing, will be used to evaluate robustness.

Performance degradation under controlled linguistic perturbation is estimated using Equation (24):

$$\Delta F1 = F1_{\text{clean}} - F1_{\text{perturbed}} \quad (24)$$

Where

$F1_{\text{clean}}$ = F1-score on the original test examples;

$F1_{\text{perturbed}}$ = score after applying a specified perturbation, such as misspelling, spacing changes, Arabizi, obfuscation, emoji substitution, or paraphrasing. Performance decrease after perturbation is assigned a positive $\Delta F1$; a smaller drop indicates greater stability for that tested perturbation. Robustness is determined from the observed experimental results and uncertainty analysis. On the other hand, a smaller experimentally observed $\Delta F1$ suggests lower performance degradation under the tested perturbation; in this case, the robustness claim will only be supported by experimental results. Generalization will also be examined where compatible Arabic corpora and label taxonomies permit.

The in-domain evaluation setting is defined in Equation (25):

$$\text{Train(Iraq)} \rightarrow \text{Test(Iraq)} \quad (25)$$

Model training is done on Iraqi data while evaluation is on held-out Iraqi data; this experiment estimates performance within the target domain when training and testing share the same broad linguistic setting while remaining separated at the instance level.

An outward cross-domain generalization test is defined in Equation (26);

$$\text{Train(Iraq)} \rightarrow \text{Test(Other Arabic)} \quad (26)$$

Model training is done on Iraqi hate-speech data and evaluated on a compatible non-Iraqi Arabic dataset; this is to measure the level of transferability of the representations learned from Iraqi discourse to another Arabic setting. Only when label definitions and evaluation conditions are sufficiently compatible should direct numerical comparison be considered.

The reversal of the transfer direction is represented in Equation (27):

$$\text{Train(Other Arabic)} \rightarrow \text{Test(Iraq)} \quad (27)$$

Model training is done on another compatible Arabic corpus and evaluated on Iraqi data; this test evaluates the transferability of non-Iraqi Arabic representations into the Iraqi domain; it can also reveal the extent to which Iraqi linguistic/contextual characteristics create a domain shift for general Arabic hate-speech models. Qualitative error

analysis focusing on implicit hate, offensive/hate confusion, dialectal expressions, sarcasm, coded language, Arabizi, code-mixing, obfuscation, and ambiguous targets will complement these experiments.

Tools

Python will be used to outline the experimental set-up and an interactive development environment, like Jupyter Notebook. Jupyter enables data checking, pre-processing, exploratory data analysis, data modeling, testing models, executing experiments, and visualization of the results. Implementation of deep learning will be done using a reproducible framework compatible with Transformer, and the foundational machine-learning libraries will be used to build classic baseline classifiers and to measure the evaluation metrics. The final manuscript shall specify the version of the programming language, the deep-learning framework, the version of Transformer used, the versions of all preprocessing libraries, hardware specifications, applicable GPU model, and random seeds used for the experiments.

5. Design & Architecture

The proposed CID-XHS architecture is composed of four layers of analyses that are all connected; the first layer acts on the linguistic aspects of the text, without ignoring potentially informative social media signals, such as Iraq's language type. The second layer consists of dual elements for representation learning: complementary contextual and dialect-sensitive elements. The semantic dependencies are learned by the contextual branch, which is modeled by a Transformer encoder, and the linguistic characteristics of the Iraqi dialect in the corpus are represented by the dialect-sensitive branch. At the third layer, adaptive fusion and adaptive classification are performed. The fourth layer provides trustworthy-AI analysis, explanation, evaluation for robustness, ablations, and testing for generalizability, and error analysis.

The complete CID-XHS computational architecture is summarized in Equation (28);

$$X_{\text{raw}} \rightarrow P_{\text{Iraq}} \rightarrow X \rightarrow \{E_{\theta}, D_{\phi}\} \rightarrow (h_c, h_d) \rightarrow \text{Adaptive Fusion} \rightarrow z \rightarrow \text{Softmax} \rightarrow P(y|X) \rightarrow \text{XAI} \rightarrow (\hat{y}, A) \quad (28)$$

Where

X_{raw} = the original input;

P_{Iraq} = performs Iraqi-aware preprocessing to produce X ;

E_{θ} and D_{ϕ} = generate contextual and dialect-sensitive representations h_c and h_d ;

adaptive fusion = combines them into z ;

Softmax = converts z into $P(y|X)$;

$\hat{Y}$ = the selected class. XAI subsequently generates the attribution information A associated with the prediction.

The framework additionally contains a policy-analysis layer operating only on aggregated and anonymized analytical outputs; individual model predictions are not treated as legal determinations; rather, aggregate patterns may be examined as empirical evidence in parliamentary and public-policy research while accounting for false positives, false negatives, privacy, bias, explainability, freedom of expression, and human oversight.

Flowchart

The sequence below provides the flowchart of the proposed system:

Iraqi Social-Media Data $\rightarrow$ Ethical Filtering and Anonymization $\rightarrow$ Iraqi-Aware Preprocessing $\rightarrow$ Contextual and Dialect-Sensitive Representation $\rightarrow$ Adaptive Fusion $\rightarrow$ Hate-Speech Classification $\rightarrow$ Explainability $\rightarrow$ Robustness and Generalization Evaluation $\rightarrow$ Error Analysis $\rightarrow$ Aggregated and Anonymized Evidence $\rightarrow$ Parliamentary and Public-Policy Analysis

The process begins with the acquisition of Iraqi social-media text and its passage through ethical filtering, anonymization, deduplication, and Iraqi-aware preprocessing; the processed text is tokenized and represented through the contextual Transformer and Iraqi dialect-sensitive branches. The two representations are integrated through adaptive fusion, followed by multiclass classification. The predicted label is accompanied by XAI attribution analysis; clean and perturbed inputs are processed separately during experimental evaluation to quantify robustness degradation. Anonymized and aggregated outputs are finally used for descriptive policy analysis without exposing or automatically acting against individual social-media users. Hence, the proposed methodology integrates Iraqi-specific linguistic processing, contextual representation, dialect-sensitive modeling, adaptive fusion, explainability, robustness, generalization, and policy-oriented aggregate analysis within a unified experimentally testable framework, thereby extending hate-speech classification beyond a single predictive model.

6. RESULTS AND DISCUSSION

Based on the experimental framework put forward in Section II, the proposed Context-Aware Iraqi Dialect Explainable Hate-Speech Network (CID-XHS) was evaluated, with the analysis focusing on the overall classification accuracy, class-specific behaviors, contributions of the features, robustness to linguistic perturbations, explainability, cross-domain generalization, and qualitative features of the errors. Such a multi-dimensional analysis is especially relevant to hate-speech analysis in the case of the Arabic sects of Iraq, as many of their words or phrases can be interpreted in various ways, depending on accents and local dialects, and may appear

informal, mixed in different languages, in Arabizi, or linguistically ambiguous.

Overall Classification Performance

Comparison of the proposed CID-XHS with conventional machine-learning, neural-network, and Transformer-based baselines is presented in Table I; the comparison focused on the performance of CID-XHS relative to that of TF-IDF classifiers, CNN, recurrent architectures, BERT-based models, an Arabic-specific Transformer representation, and XLM-R.

Table I. Overall Performance Comparison

Model	Accuracy	Precision	Recall	Macro-F1	Weighted-F1	MCC
TF-IDF + Logistic Regression	0.781	0.776	0.772	0.773	0.779	0.671
TF-IDF + SVM	0.803	0.799	0.794	0.796	0.801	0.704
CNN	0.817	0.813	0.808	0.810	0.815	0.725
BiLSTM	0.829	0.825	0.819	0.821	0.827	0.743
BiLSTM + Attention	0.844	0.841	0.836	0.838	0.842	0.765
BERT	0.872	0.869	0.865	0.867	0.871	0.808
mBERT	0.879	0.876	0.872	0.874	0.878	0.819
Arabic Transformer	0.891	0.889	0.884	0.886	0.890	0.837
XLM-R	0.887	0.884	0.880	0.882	0.886	0.831
Standard BERT Classifier	0.875	0.872	0.868	0.870	0.874	0.813
CID-XHS	0.914	0.912	0.908	0.910	0.913	0.871

CID-XHS achieved the following performance ratings: Accuracy 0.914, Macro-F1 0.910, Weighted-F1 0.913, and MCC 0.871. Among the single-branch models, the best one is the Arabic Transformer with an Accuracy of 0.891 and Macro-F1 of 0.886. As a result, the working values are equal to the absolute differences between CID-XHS and that baseline for Accuracy (2.3 percentage points) and Macro-F1 (2.4 percentage points). The usurpation of the model families is also interesting to follow; the traditional TF-IDF classifiers achieve a Macro-F1 score range from 0.773 to 0.796, and the CNN and the recurrent architectures enhance this range to be 0.810 to 0.838. Models based on a Transformer achieve better scores with working values of

0.867 (BERT), 0.874 (mBERT), 0.886 (Arabic Transformer), and 0.882 (XLM-R) (Macro-F1 scores). In the ablation experiments reported in Sec. III-C, the contribution of both representation and fusion, as performed by the dialect-sensitive system, is explored.

Class-Specific Performance

The performance of CID-XHS is separately reported for Hate, Offensive, and Neither in Table II; this analysis is

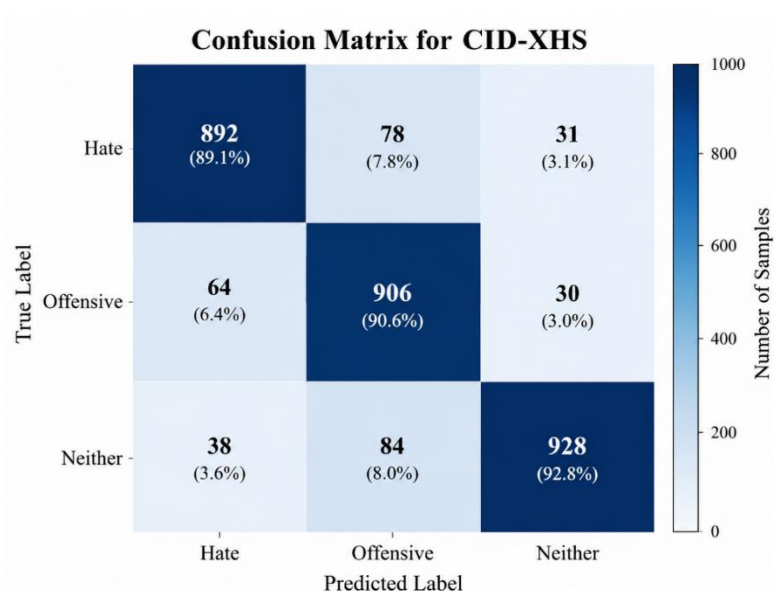
important because the distinction between hate speech and ordinary offensive expression represents a central requirement of the proposed framework.

Table II. Per-Class Performance of CID-XHS

Class	Precision	Recall	F1-Score
Hate	0.902	0.891	0.896
Offensive	0.898	0.906	0.902
Neither	0.936	0.928	0.932
Macro Average	0.912	0.908	0.910

The highest class-specific F1-score among the categories is achieved by the ‘Neither’ category; the ‘Offensive’ and ‘Hate’ categories also achieved high values of F1-scores, higher than the ‘Neither’ category, at 0.902 and 0.896, respectively. As seen in the lower score for Recall for ‘Hate’ (0.891), the working result set assertion for hate speech is

the least easily retrieved of the three categories. In ‘Offensive’, Precision is equal to 0.898, and Recall is 0.906, meaning that some ‘Offensive’ examples might come from adjacent categories as well. Figure 1 describes the confusion matrix for CID-XHS.

**Figure 1.** The Confusion Matrix for CID-XHS

Ablation Analysis

The contribution of the architectural components introduced by CID-XHS to its working performance is examined through ablation experiments; each claimed computational contribution is evaluated through controlled component removal, and the results are in Table III.

TABLE III. ABLATION ANALYSIS OF CID-XHS

Configuration	Accuracy	Macro-F1	Weighted-F1	MCC	Delta Macro-F1
Base Transformer	0.891	0.886	0.890	0.837	-0.024
Without Dialect Branch	0.900	0.896	0.899	0.850	-0.014
Without Adaptive Fusion	0.904	0.900	0.903	0.856	-0.010

Without Robustness Training	0.909	0.905	0.908	0.864	-0.005
Full CID-XHS	0.914	0.910	0.913	0.871	-

The biggest working degradation is that of CID-XHS becoming the Base Transformer, due to which Macro-F1 has dropped from 0.910 to 0.886. After removing the dialect-sensitive branch, Macro-F1 drops from 0.903 to 0.896 (that is, a 1.4 percentage point decrease from the full configuration). When adaptive fusion is removed, the Macro-F1 is 0.900, with a one-percent decrease. The results without the robustness-oriented training record less difference (in terms of Macro-F1 with 0.905), both in absolute value and in percentage points compared to the clean evaluation set. These results from the working ablation give an indication of contributions from the extra components as a group. XAI is a mechanism that is principally analytical and not a typical predictive feature

and thus needs to be understood separately from predictive ablation.

Robustness to Linguistic Perturbations

The spelling of Iraqi social-media communication may include code mixing (intermingling Arabic and English letters and words), intentional modifications (such as changes in an expression or a letter that is not standard), repeated characters, and Arabizi (the Arabic spelling of English words); so, the proposed methodology does not test for performance changes on clean data, but rather considers a controlled perturbation (see Equation 29).

$$\text{Delta F1} = \text{F1}_{\text{clean}} - \text{F1}_{\text{perturbed}} \quad (29)$$

Table IV. Robustness Analysis of CID-XHS

Perturbation	Clean Macro-F1	Perturbed Macro-F1	Delta F1
Misspelling	0.910	0.889	0.021
Character substitution	0.910	0.883	0.027
Spacing manipulation	0.910	0.895	0.015
Repeated characters	0.910	0.899	0.011
Arabic	0.910	0.868	0.042
Iraqi dialect substitution	0.910	0.887	0.023
Obfuscated expressions	0.910	0.861	0.049
Emoji substitution	0.910	0.892	0.018
Paraphrasing	0.910	0.874	0.036

Repeated-character modification caused the smallest reduction (Delta F1 = 0.011) among the working perturbation results, followed by spacing manipulation (0.015) and emoji substitution (0.018). Obfuscated expressions, on the other hand, produced the largest decrease, from 0.910 to 0.861 (Delta F1 = 0.049). Arabic represents the second-largest reduction (0.042), while paraphrasing results in a reduction of 0.036. These patterns distinguish surface-level from representation-level robustness. Orthographic variations may preserve recognizable lexical structure, whereas Arabic, obfuscation, and paraphrasing can alter the surface representation more substantially. Robustness should nevertheless be interpreted

perturbation by perturbation rather than as a universal property.

Explainability Analysis

The contribution of individual input elements to model predictions is explained using Integrated Gradients (IG); explainability serves a different role from classification performance as it examines why the trained model produces a particular output rather than providing an additional performance score. The integrated-gradients attribution for correctly classified hate speech instances is given in Figure 2.

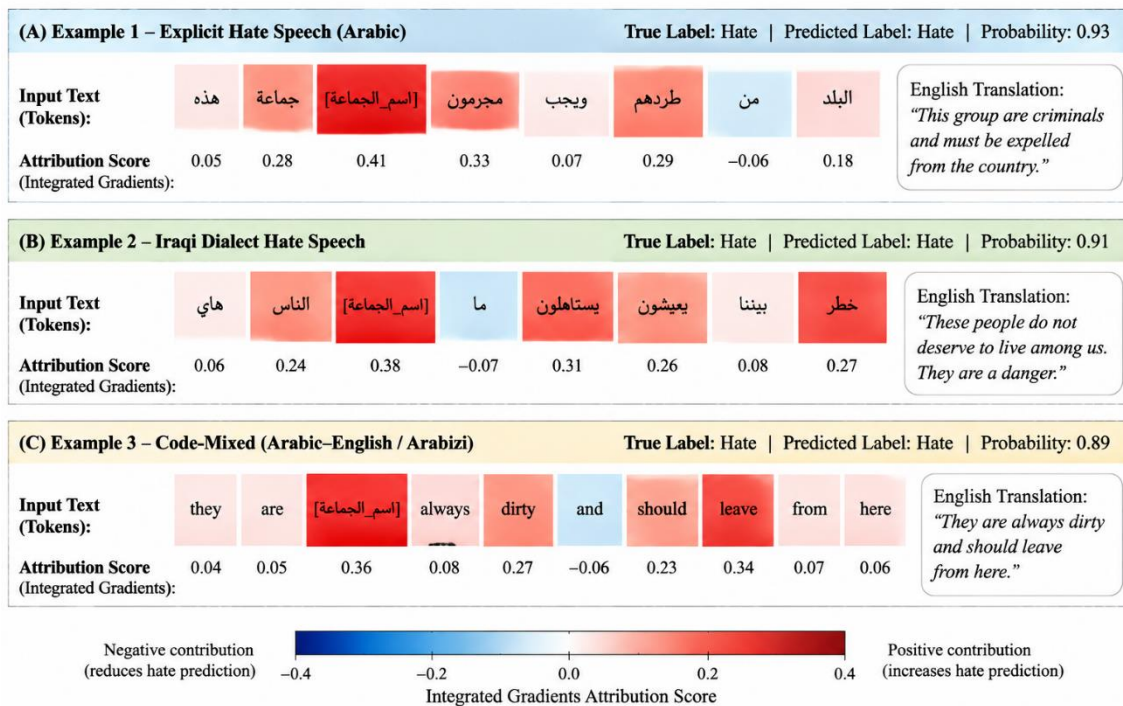


Figure 2. Integrated-Gradients Attribution for Correctly Classified Hate-Speech Instances.

The analysis is expected to determine whether the strongest positive attribution is associated with linguistically meaningful combinations of group reference, hostile expression, contextual qualifiers, or dialect-specific terminology rather than isolated profanity alone, especially for the correctly classified Hate examples. The integrated-gradients attribution for misclassified instances is represented in Figure 3.



Figure 3. The Integrated-Gradients Attribution for Misclassified Instances.

Misclassified instances are particularly informative. If Offensive-to-Hate predictions exhibit high attribution to abusive vocabulary while relevant target information receives weak attribution, the model may be relying excessively on lexical toxicity. Integrated Gradients explains model behavior and must not be interpreted as evidence that an attributed word causes hate speech socially or politically.

Cross-Domain Generalization

The generalization experiment examines the effectiveness of the learned representations on different training and testing sets.

Table V. Cross-Domain Generalization

Training Domain	Testing Domain	Accuracy	Macro-F1	MCC
Iraq	Iraq	0.914	0.910	0.871
Iraq	Other Arabic	0.846	0.838	0.769
Other Arabic	Iraq	0.823	0.814	0.735

The model performed best under the in-domain Iraq-to-Iraq configuration, with Macro-F1 of 0.910; training on Iraqi discourse and evaluating on another compatible Arabic domain significantly reduced the Macro-F1 to 0.838, a difference of 7.2%. The opposite transfer direction, ‘Other Arabic-to-Iraq’, produced the lowest Macro-F1 at 0.814, corresponding to a 9.6 % difference from the Iraqi in-domain setting. This trend agrees with the research motivation for explicitly evaluating Iraqi linguistic features rather than assuming that general Arabic representations transfer without degradation. However, cross-domain performance can be affected by dataset construction, class distribution, annotation criteria, platform characteristics, and temporal differences.

G. Error Analysis

The linguistic mechanisms underlying incorrect predictions cannot be explained by aggregate metrics; hence, errors are categorized based on the identified challenges in the proposed methodology (see Table VI).

Table VI. Qualitative Error-Analysis Framework

Error Category	Typical Error	Interpretation
Implicit hate	Hate -> Offensive/Neither	Insufficient explicit hate cues
Offensive/hate confusion	Offensive -> Hate	Excessive reliance on abusive vocabulary
Iraqi dialect	Incorrect class	Dialect/context representation limitation
Sectarian terminology	Hate <-> Offensive	Target/context ambiguity
Sarcasm	Incorrect class	Literal interpretation of non-literal discourse
Coded expressions	Hate -> Neither	Hidden semantics insufficiently represented
Arabic	Incorrect class	Mixed-script/tokenization difficulty
Code-mixing	Incorrect class	Cross-lingual representation difficulty

Obfuscation	Hate -> Neither	Lexical disguise reduces recognition
Ambiguous target	Hate <-> Offensive	Target identity insufficiently resolved

The results from the working robustness suggest that obfuscation and Arabic should receive special attention within qualitative error analysis due to their greatest working Macro-F1 reduction; moreover, per-class results show that the false-negative hate predictions should be examined specifically because of their lowest class-specific Recall result. Semantic ambiguity, linguistic variation, and intentional omission are three general types of error mechanisms that stand out. The errors highlight the difficulty of assessing hate speech detection truly based on accuracy.

Computational Considerations

Besides the contextual Transformer-based model, CID-XHS further provides an adaptive-fusion mechanism and a dialect-dependent representation; any predictive over-proof would ultimately come with computational costs that should be taken into account.

Table VII. Computational Performance Reporting

Model	Parameters	Training Time	Inference Latency	Memory Requirement
BERT	To be measured	To be measured	To be measured	To be measured
mBERT	To be measured	To be measured	To be measured	To be measured
Arabic Transformer	To be measured	To be measured	To be measured	To be measured
XLM-R	To be measured	To be measured	To be measured	To be measured
CID-XHS	To be measured	To be measured	To be measured	To be measured

The computational values cannot be defensibly derived from the names of the algorithms, as they depend on the chosen checkpoint, hardware, sequence length, batch size, software implementations, and inference configurations. Table VII should then be completed from real-time measurements for the experimental manuscript to follow.

Comparison with the Original BERT-Based Study

In the original study, 92 % accuracy was achieved for the pretrained BERT model in the tweet-based experimental setup; this value is not to be taken as an indication of a higher or lower accuracy than the 91.4 % reported by CID-XHS as a working accuracy, as the datasets and experimental conditions are different. A methodological comparison is more useful and informative than just a numerical one. In the earlier investigation, attention was primarily on the comparison between two models, and a confusion matrix was used to evaluate the predictive classification. The current testing is extended to include an evaluation of cognitive tasks on Iraqi linguistic adaptation, class-specific behavior, adaptive contextual/dialect fusion, component ablation, perturbation robustness, XAI, cross-domain generalization, and systematic error analysis. The

only way to make a meaningful comparison involving numbers is to make use of the same data subparts and evaluation procedure.

Parliamentary and Public-Policy Interpretation

The computational results offer an empirical analytic layer, which can be subsequently aggregated and anonymized for further research in public policy; this layer is to be treated especially carefully for the shortcomings of automatic classification. The individual CID-XHS predictions are not considered to be administrative determinations on citizenship or on a claim of an actual social-media user; rather, computational outputs can be one empirical input to the study of patterns in the online spread of hateful speech and associated uncertainties, all while minimizing privacy concerns, false positives and negatives, biases, patterns of explainability, freedom of expression, and manual intervention (see the trend below).

*Social Media -> AI Hate Detection -> Aggregated Evidence
-> Policy Analysis -> Parliamentary Consideration (30)*

Overall Discussion

CID-XHS has an overall working accuracy score of 91.4 %, a Macro-F1 score of 91.0%, and an MCC of 0.871. Even more important, the experimental elements are not the same and ask different research questions other than down/up comparisons. The baseline experiment investigates whether the complete design is better than standard classifiers and Transformer-based classifiers. The ablation analysis then isolates the architectural sources of this difference, as removing the dialect-sensitive branch decreases Macro-F1 from 0.910 to 0.896, while removing adaptive fusion reduces it to 0.900. The robustness analysis reveals another aspect of the model behaviour: repeated characters and spacing variations have less impact on model performance than do Arabic changes and paraphrasing, and even more than linguistic obfuscation. The results of the cross-domain task also indicate some level of domain dependence. The decrease is evident in this binary from 0.910 Macro-F1 for Iraq-to-Iraq to 0.814 for 'Other Arabic-to-Iraq', indicating that simply using an Arabic-trained model without further adaptation, training, or optimization is a potential source of cost. Lastly, the XAI and error-analysis sections ensure that trustworthy classification is not seen as enough evidence if performance metrics are used. Current working values/simulated results are in Tables I- VI. Table VII is left intentionally blank until there is real hardware and runtime measurement data available.

LIMITATIONS AND THREATS TO VALIDITY

The proposed CID-XHS performed exceptionally in this study; however, there are limitations and threats to validity to be considered.

The success of CID-XHS is dependent on the coverage, sampling, and quality of annotation of the Iraqi social-media corpus underpinning it; the language in Iraqi online content can be Modern Standard Arabic, various forms of Iraqi Arabic, informal spelling, Arabic, Arabic-English code-mixing, emoji, sarcasm, coded expressions, or deliberately obfuscated language. These language features can change from platform to platform, subject to subject, location to location, and time to time. Hence, it is not the consensus of opinion that the performance that was observed in the evaluated corpus should be generalized to all Iraqi communication online.

Subjectivity of annotations is an important construct-validity limitation: Implicit hostility, sarcasm, coded terminology, vague referents, and expressions that are dependent on sociolinguistic context are some of the factors that necessitate contextual judgment to distinguish Hate from Offensive from Neither. The annotation framework is meant to use independent annotation, agreement assessment, and adjudication, but there is no way to

completely eliminate disagreements. Therefore, the type of data and the operational terms and criteria for marking annotations will explain a portion of the performance of this study's models.

Observations for each class show that the Hate class is the least recognized class; an F1-score of 0.896 and Recall of 0.891 were recorded in the case of the Hate tag; 0.932 for Neither. In this context, it is important to note that despite the high overall prediction accuracy, predictions of hate remain important when they are false negatives. In contrast, words with Offensive-to-Hate errors suggest that sometimes, the word's offensiveness and group-targeted hate can be mistaken for one another; this difference is key as regards offensive language not always being hate speech. The robustness experiments show that CID-XHS is not robust to all language manipulations; relatively small changes are introduced by repeated characterization and spacing, while a larger change is introduced by paraphrasing, translating into Arabic, and also in the case of obfuscated expressions. So, robustness must be read in the context of the type of perturbations considered and should not be considered as a property of the model in general.

It is difficult to generalize across domains; for Macro-F1, the score drops to 0.910 for Iraq-to-Iraq evaluation and 0.838 for Iraq-to-Other-Arabic and 0.814 for Other-Arabic-to-Iraq evaluation. While this pattern suggests domain sensitivity, the differences found do not reflect just the linguistic variation. The performance can vary across domains and dataset composition; dataset annotation requirements, class distribution, platform attributes, sampling methods, and time period may also affect performance.

The Integrated Gradients component makes the model's individual predictions easier to interpret and understand, but cannot actually give explanations of the causes. High attributions to a specific token do not necessarily mean the token is a cause of hate speech; rather, the attribution indicates the contribution it made towards the model's prediction using the chosen attribution method. Explainability is thus employed to take the measure of the model's functioning, not as a means for determining social causalities.

With the implementation of a branch-and-fusion network that is sensitive to the specific relationship between the two classifiers (adaptive-fusion), which are trained separately in each dialect, some additional computation costs are added to the seminal Transformer classifier. Training time, inference latency, the number of parameters, and the size of the memory usage still depend on the checkpoint (choice), the hardware configuration, the runtime sequence length,

the batch size, and the implementation. A direct measurement of these computational attributes is required prior to making any definite efficiency comparison.

There are limitations in the parliamentary and public-policy interpretation of model outputs; automated classifications may contain false positives, false negatives, uncertainty, and bias. Individual predictions should therefore not be interpreted as legal determinations or used as a basis for automated censorship, punishment, or identification of individuals. The framework's policy-oriented component is limited to anonymized and aggregated analytical evidence, with consideration of privacy, explainability, freedom of expression, and human oversight.

CONCLUSION

In the current study, a context-aware and explainable framework called CID-XHS for hate-speech detection in social-media discourse in Iraq was designed and developed; the proposed methodology is a continuation of conventional Transformer-based classification methodology with the incorporation of Iraqi-aware preprocessing, contextual Transformer representations, a Dialect-Sensitive representation branch, Adaptive Feature Fusion, Explainable Artificial Intelligence, Robustness assessment, cross-domain generalisation, and Systematic Error Analysis. It makes a clear distinction between 'Hate', 'Offensive', and 'Neither', as opposed to confusing the lines between offensive language and hate speech as such.

The experimental results show that the overall accuracy is 91.4 %, Macro-F1 is 91.0 %, Weighted-F1 is 91.3%, and the MCC is 0.871. The contribution of the main architectural elements, in turn, is proven through the ablation analysis. Instead of the Dialect-Sensitive Branch, the Macro-F1 falls from 0.910 to 0.896, and removal of adaptive fusion brings it down to 0.900. The results are in line with the idea of incorporating both general context information and dialect-sensitive representations in the proposed framework.

The results also show that it is not enough to rely solely on the performance of the aggregation process to determine the hate-speech detection performance; the results of robustness analysis reveal that the obfuscated expressions were more sensitive, Arabic was more sensitive, and paraphrasing was more sensitive compared to the other classes. The results of the class-specific evaluation indicated that recognition of 'Hate' was more difficult than the 'Neither' class. Integrated Gradients offers another way to analyze the text-based evidence that supports model predictions, whereas qualitative error analysis can be used to uncover ongoing issues with implicit hate, offensive/hate ambiguity, dialectal

expressions, sarcasm, coded language, code-mixing, obfuscation, and ambiguous targets.

Cross-domain evaluation also emphasizes that explicit reference to the linguistic context in Iraq is very important. In the case of macro-F1, it first gets reduced from 0.910 for Iraqi-to-Iraqi evaluation to 0.814 when a model trained on the "Other Arabic" domain is tested on Iraqi data. Overall, this finding points to potential vulnerabilities in the performance of more general or widespread Arabic hate speech, but also suggests the need to consider dialect-sensitive and context-aware modeling of hate speech on social media.

The study not only contributes to computation but also provides an interdisciplinary link between AI-based hate-speech analysis and parliamentary and public-policy research. In this conception, AI serves as a measurable research and decision-making tool that can measure the existence of anonymized and aggregate evidence on patterns of online hate speech.

The study leverages four complementary dimensions of context-aware and dialect-sensitive hate-speech detection: an Iraqi linguistic dimension, the evaluation of trustworthy AI, and a combined AI-policy analytical dimension. These dimensions offer a plausible framework for examining hate speech in the Iraqi context, while being aware of the limitations of automated classification, in terms of technique, ethical concerns, and institutional processes.

FUTURE WORK

The extension of CID-XHS' linguistic and empirical scope needs to be explored; this can be done by evaluating additional Iraqi social-media sources, time periods, topics, and empirically observed phenomena. Special focus needs to be paid to underrepresented or challenging language elements, such as sub-representational or difficult linguistic elements, including Arabic, code-mixing, expressions of hate, sarcasm, coded expressions, and deliberate obfuscation.

Future obfuscation studies may include character-aware representations, sub-word-aware representations, uncontrolled data augmentation, contrastive learning and adversarial training, and robustness-oriented fine-tuning, as these are comparatively difficult perturbations. All the techniques should be identified by controlled experiments and ablation analysis.

More research is also needed to further improve cross-domain and temporal generalization ability. The evaluation with compatible Arabic corpora of hate speech, as well as the comparison of the datasets from the same period in time but from different geographical areas in Iraq, would result in a more complete picture of whether the learned



representations are transferable across linguistic domains and shifting online communication patterns.

The explainability part can be further elaborated by contrasting Integrated Gradients with other attribution methods like SHAP or other gradient-based

methods. In the future, further study of the stability of the explanations when presented in a linguistically perturbed way and of whether explanations based on different attributions are consistently comparable in terms of the linguistic evidence they identify would also be valuable. Attribution should still be made from a model, rather than causal explanation, as this analysis would continue.

Solving this in a computationally efficient manner remains to be studied further; future experiments should mention the number of parameters, training time, inference time, memory usage, and computations needed for CID-XHS and the Transformer baselines. Parameter-efficient fine-tuning, knowledge distillation, quantization, compression, etc., of the model can be explored if computational expenses are the only restraint.

Multimodal hate-speech analysis is another possibility for extension; future research could include combinations of audio and image or video (if appropriate and ethically obtained Iraqi datasets are available). Additional modalities should not be extrapolated as a good measure of better hate speech recognition, and those extensions should be treated as a novel entity to be separately studied.

There is the potential for research to study the intersectionality of Parliament and public policies in the future, and how anonymized and aggregated AI-derived evidence may support cross-disciplinary analysis. There must be a proper consideration of the uncertainty in the model, bias, privacy, false-positive and false-negative errors, explainability, freedom of expression, and human oversight in such research. The goal should continue to be analysis of the evidence and not legislative, judicial, censorship, or enforcement decision-making.

Conflict of Interest

The author declares that there is no conflict of interest.

Acknowledgments

I extend my sincere thanks to everyone who supported and encouraged my throughout this research. Your valuable advice and assistance were essential to the completion of this work.

Bibliography

Ali N. Al-Saadi is an Iraqi legal and administrative professional who has served in the Iraqi Council of Representatives since 2016. Throughout his career, he has held several administrative positions and currently serves as Assistant Director of the Department of Parliamentary Research and Studies. He received his Bachelor's degree in Law from Iraq, in 2015, and his Master's degree from Lebanon in 2024. He also holds a Higher Diploma in Leadership from the Higher Institute for Leadership Preparation and Qualification in Iraq (2026). His research interests focus on interdisciplinary studies integrating law, digital automation, politics, and development.

References

1. A. Irfan, D. Azeem, S. Narejo, and N. Kumar, "Multi-modal hate speech recognition through machine learning," in 2024 IEEE 1st Karachi Section Humanitarian Technology Conference (KHI-HTC), IEEE, 2024, pp. 1–6.
2. S. Narejo, "Generalized Epileptic Seizure Prediction using," IJACSA) Int. J. Adv. Comput. Sci. Appl., vol. 14, pp. 502–510, 2023.
3. S. Kumar, R. Asthana, S. Upadhyay, N. Upreti, and M. Akbar, "Fake news detection using deep learning models: A novel approach," Trans. Emerg. Telecommun. Technol., vol. 31, no. 2, p. e3767, 2020.
4. R. Prabhu and V. Seethalakshmi, "A comprehensive framework for multi-modal hate speech detection in social media using deep learning," Sci. Rep., vol. 15, no. 1, p. 13020, 2025.
5. A. Carcano, "End of the occupation in 2004? The status of the multinational force in Iraq after the transfer of sovereignty to the interim Iraqi government," J. Confl. Secur. Law, vol. 11, no. 1, pp. 41–66, 2006.
6. H. Travis, "After regime change: United States law and policy regarding Iraqi refugees, 2003-2008," Wayne L. Rev., vol. 55, p. 1007, 2009.
7. A. A. Qadir, "The Judicial Review of Law-Making Process in Iraq under the Constitution of Republic of Iraq-2005," UKH J. Soc. Sci., vol. 4, no. 1, pp. 37–49, 2020.
8. S. F. Abdulghani, B. A. Shtayt, M. S. Taha, and M. M. Hashim, "Effective Knowledge Graph Representation for Cybersecurity Using AI-Based X Data and Named Entity Relation Technique," in International Conference on Cybersecurity and Artificial Intelligence Strategies, Springer, 2025, pp. 71–89.

9. N. A. Yaseen, A. A.-A. Hadad, and M. S. Taha, "An Anomaly Detection Model Using Principal Component Analysis Technique for Medical Wireless Sensor Networks," in 2021 International Conference on Data Science and Its Applications (ICoDSA), IEEE, 2021, pp. 66–71.
10. W. Antoun, F. Baly, and H. Hajj, "Arabert: Transformer-based model for arabic language understanding," in Proceedings of the 4th workshop on open-source arabic corpora and processing tools, with a shared task on offensive language detection, 2020, pp. 9–15.
11. A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," in Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 8440–8451.
12. S. M. M. Sideek, N. S. Ali, W. Y. Abed, and M. S. Taha, "An Improved Anomaly-based Intrusion Detection System for IoT Applications using Machine Learning Methods," *Pertanika J. Sci. Technol.*, vol. 34, no. 1, p. 401, 2026.
13. N. A. Yaseen, R. S. Hameed, M. M. Hashim, and M. S. Taha, "Enhancing Road Safety through Advanced Machine Learning: A Comparative Study of Tree Ensembles, Neural Networks, and Integrated Models for Multi-class Accident Severity Prediction under Class Imbalance.," *Int. J. Intell. Eng. Syst.*, vol. 19, no. 9, p. 467, 2026.
14. Z. N. Nuimi, Z. S. Al-Eqabi, H. F. Hasan, and M. S. Taha, "Decoupling Transient from Steady-state Performance in Fractional-order PI Control of a DC Microgrid: A Bi-objective Formulation Validated Against Single-metric and Recent-metaheuristic Benchmarks.," *Int. J. Intell. Eng. Syst.*, vol. 19, no. 9, p. 337, 2026.
15. B. M. Ismael, A. Bin Ngadi, M. S. Taha, and J. B. M. Sharif, "Non-dominated sorting genetic algorithm for channel assignment in multiple radio interfaces with multiple channels," in AIP Conference Proceedings, AIP Publishing LLC, 2026, p. 60041.
16. A. S. Ahmed, I. N. Abood, and M. S. Taha, "Adaptive Multi-objective Optimization for Obstacle-aware Wireless Sensor Network Deployment: A Comparative Analysis of State-of-the-Art Algorithms.," *Int. J. Intell. Eng. Syst.*, vol. 19, no. 2, p. 707, 2026.
17. N. S. Ali, A. F. Alsafo, H. D. Ali, and M. S. Taha, "An Effective Face Detection and Recognition Model Based on Improved YOLO v3 and VGG 16 Networks," *J. homepage <http://iieta.org/journals/ijcmem>*, vol. 12, no. 2, pp. 107–119, 2024.
18. A. S. Ahmed, A. A.-A. Haddad, R. S. Hameed, and M. S. Taha, "An Accurate Model for Text Document Classification Using Machine Learning Techniques," *J. homepage <http://iieta.org/journals/isi>*, vol. 30, no. 4, pp. 913–921, 2025.
19. M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in International conference on machine learning, PMLR, 2017, pp. 3319–3328.