Available at https://www.iasj.net/iasj
Iraqi Academic Scientific Journals
Journal homepage:
https://journals.uokerbala.edu.iq/index.php/UOKJ

Research Article

The Importance of LIME Interpretations and Training Curve Analysis in valuating Hate Speech Detection Models

Safaa Hameed Kareem [1] , Asia Mahdi Naser alzubaidi [2]
[1,2]college of computer science and information technology university of kerbala ,Kerbala, Iraq

Article Info
Article history:
Received 6-11-2025
Received in revised form 15-12-2025
Accepted 22-1-2026
Available online 31 -3-2026
Keywords: Hate Speech detection, Interpreter LIME, Training Curve.

Abstract:

The necessity of creating efficient techniques for identifying offensive language, especially hate speech, has increased due to the quick expansion of user-generated material on social media platforms. Even though machine learning and deep learning models have made significant strides in this area, accuracy and related quantitative metrics are frequently the only criteria used to evaluate them. Despite their usefulness, these metrics fall short in capturing the intricacies of learning dynamics or the interpretability of model predictions. By including training curve analysis and Local Interpretable Model-agnostic Explanations (LIME) into the assessment of hate speech recognition models, with an emphasis on the Iraqi Arabic dialect, this study overcomes this problem.

We experimented with three different methods: first, using training FastText and BiLSTM separately, second, BiLSTM with FastText embeddings, and third combining FastText embeddings with lexicon-based features. A closer examination of the second method showed significant subtleties, even though it had the greatest reported accuracy of 97.3%. While LIME interpretations offered insight into the linguistic elements underlying classifications, training curves revealed problems with overfitting and stability. These revelations highlighted the significance of interpretability and training dynamics by exposing misclassifications that accuracy alone was unable to account for.

The results emphasize how important it is to implement thorough assessment frameworks that strike a balance between qualitative comprehension and quantitative performance. These methods guarantee the creation of hate speech detection models for Arabic dialects that are more resilient, trustworthy, and context-sensitive.

Corresponding Author E-mail: safaa.h@s.uokerbala.edu.iq; asia.m@uokerbala.edu.iq
Peer review under responsibility of Iraqi Academic Scientific Journal and University of Kerbala.

1. Introduction:

The necessity for efficient techniques to identify and reduce harmful language, especially hate speech [1], has increased due to the quick growth of user-generated material on social media platforms [2, 3, 4]. Hate speech compromises the integrity of digital communication platforms in addition to posing a threat to the welfare of persons and communities [5]. Even while the use of machine learning and deep learning models has led to significant advancements, the assessment of these models is still not well studied, despite its crucial significance. Traditionally, accuracy and associated quantitative measures like precision, recall, and F1-score have been used to evaluate performance. Although these metrics offer helpful standards, they frequently fall short of capturing the more complex dynamics of learning or the interpretability of model choices, two critical aspects of implementing models in delicate fields, such as hate speech detection.

Accuracy-based evaluation frequently has the drawback of being unable to identify problems like instability during training, underfitting, or overfitting. Two models may have similar or even better accuracy scores, but their ability to generalize to new data is essentially different. A useful addition is training curve analysis, which shows the learning trajectory over epochs and indicates if a model overfits the training set, converges smoothly, or misses underlying patterns. These kinds of insights are necessary to guarantee that the high accuracy reported in real-world applications translates into consistent and dependable performance.

An efficient method for resolving these issues is Local Interpretable Model-agnostic Explanations (LIME) [6], which offers intelligible explanations of specific predictions. LIME can reveal the advantages and disadvantages of models by identifying the language signals or textual patterns that affect categorization results [7]. However, previous research has not fully integrated interpretability techniques like LIME and training curve analysis with conventional assessment measures, especially when it comes to hate speech detection. This disparity emphasizes the lack of thorough assessment systems that can strike a balance between reliability, transparency, and quantitative performance.

To investigate how evaluation metrics other than accuracy affect model performance, three methods for identifying hate speech in the Iraqi dialect were evaluated using a dataset of Iraq-related YouTube user comments:

- Training FastText and BiLSTM separately.
- Training BiLSTM model using FastText embeddings.
- Training BiLSTM model using FastText embeddings combined with lexical features.

In order to fill the research gap, this study evaluates hate speech recognition models in the Iraqi Arabic dialect by integrating training curve analysis with LIME-based interpretability. Beyond static accuracy ratings, it shows how training dynamics expose problems like instability and overfitting, and LIME offers clear insights into how the model makes decisions. When combined, these techniques combine qualitative comprehension with quantitative performance to create a thorough framework for model assessment. The findings demonstrate the need of including interpretability and learning trajectories in hate speech

detection studies, which will result in more reliable and strong computer models.

1. Related Work

Research on Modern Standard Arabic (MSA) has gained more attention lately, particularly in related domains like text categorization and sentiment analysis. Recent research has started to concentrate on Arabic dialects since each Arab nation has its unique dialect and the majority of social media users prefer to write in dialect rather than Modern Standard Arabic.

The research on Classical Arabic is reviewed below, and then the research on Arabic dialects, including the Iraqi dialect, is reviewed.

(Mousa et al. 2024) created a five-category multi-class system for inappropriate Arabic tweets. The efficacy of hybrid models in detecting objectionable language was demonstrated by their cascaded ArabicBERT–BiLSTM–RBF strategy, which used deep learning, transformers, and conventional models to reach 98.4% accuracy and F1-score [1].

(Alkhatib et al. 2024) used hybrid models, CNN, and RNN to identify instances of cyberbullying in Arabic tweets. LSTM demonstrated the efficacy of deep learning in detecting Arabic cyberbullying with a binary classification accuracy of 95.6%, while CNN led the six-class classification with 78.8% accuracy [5].

(Daouadi & colleagues 2024) [8] and (Zaghouani and Biswas 2025) [9] When it came to identifying subtle hate speech in a multilabel Arabic Twitter corpus, AraBERTv2 scored better than competing models.

Seventy thousand words from five Arabic dialects and four hate speech categories make up the ADHAR corpus (Charfi et al., 2024). Robust evaluation is made possible by its rigorous annotation and balance. Experiments revealed that models could achieve up to 90% accuracy, and that the refined AraBERT could detect hate speech with 94% accuracy and categorize it with 95% accuracy [10].

(Asiri and Saleh 2024) presented SOD, a 24,000-tweet Saudi Arabic corpus for detecting inappropriate language, such as insults, sarcasm, and hate speech. Their F1-score of 91%, which they attained by utilizing deep learning, machine learning, and AraBERT with data augmentation, emphasizes the significance of dialect-specific resources [11].

(Abutiheen et al. 2022) developed the "identity classifier" to differentiate between positive and negative Facebook comments after analyzing Arabic social media with an emphasis on the Iraqi dialect. Achieving 85.4% accuracy, it outperformed classical classifiers on the same dataset [12].

(Sabri and Abdullah, 2025) detected extremism in the Iraqi dialect (offensive and non-offensive) using machine learning using Word2Vec and FastText embeddings. With an F1 score of 0.95, the study was assessed using datasets from Twitter and Facebook [13].

The research cited above show that none of them combined interpretability analyses with thorough model evaluation measures. While some research simply took performance metrics like precision, recall, or F1-score into account, others just looked at reporting accuracy. An insufficient understanding of model behavior and reliability is offered by such methods. As this study will further show, accuracy by itself cannot be considered the

final criterion for choosing one model over another.

2. Methodology

3.1Data Collection

YouTube, which has a large collection of films about Iraqi issues, is where the dataset was gathered. Both abusive and non-abusive user comments were collected in large quantities. More than 500,000 entries made up the final corpus, which was then assembled into an Excel file and ready for preparatory operations like text normalization and data cleaning.

3.2Data Preprocessing

It is commonly known that the Iraqi dialect does not adhere to the grammatical rules of Modern Standard Arabic (MSA). Consequently, there is currently no model that can accurately handle data in the Iraqi dialect.

Removing unnecessary elements such as URLs, user mentions, HTML tags, non-text codes, tags (where they are not semantically relevant), and pointless spam comments was the first step in the data pre-processing procedure. At this stage, common issues like incomplete records and duplicate entries are also covered. Normalization is performed to normalize text representation following cleaning. This could entail standardizing spelling variations and dealing with odd letters or diacritical symbols. The database contains 120,000 records after data processing.

3.3Data Annotation

Accurate text annotation, is a basic task in natural language processing (NLP) [14,15], which is the process of giving unstructured text labels or categories (e.g., named entities or sentiment tags) to convert it into structured data. The comments were divided into four groups: Normal, Offensive language, Abusive, Hate speech (instigation to violence). Each category requires a precise definition in order to be distinguished from the others.

- The term "abusive language" describes vulgar or insulting language that is usually seen as inappropriate for usage in public settings and disapproved of by society. These include derogatory language that directly mentions sensitive body parts or is sexually explicit.
- “Offensive language” is any unpleasant conduct or statement intended to ridicule or disparage someone. For instance, using strong but non-abusive words or comparing someone to an animal are examples.
- "Hate Speech" refers to expressions that harm individuals or groups by using threats, intimidation, or identity-based targeting. This involves encouraging violence, social exclusion, and derogatory comments regarding marginalized populations.

3.4Data Balancing

Upon classifying the comments into four groups, it was discovered that the sample distribution varied unevenly throughout the groups. By oversampling the minority classes, data balancing was used to guarantee objective model training and maintain the linguistic features of the Iraqi dialect. Effective balance provides equitable learning in every class and improves generalization abilities [16]. Data balance thus becomes a crucial component of training dataset preparation

.

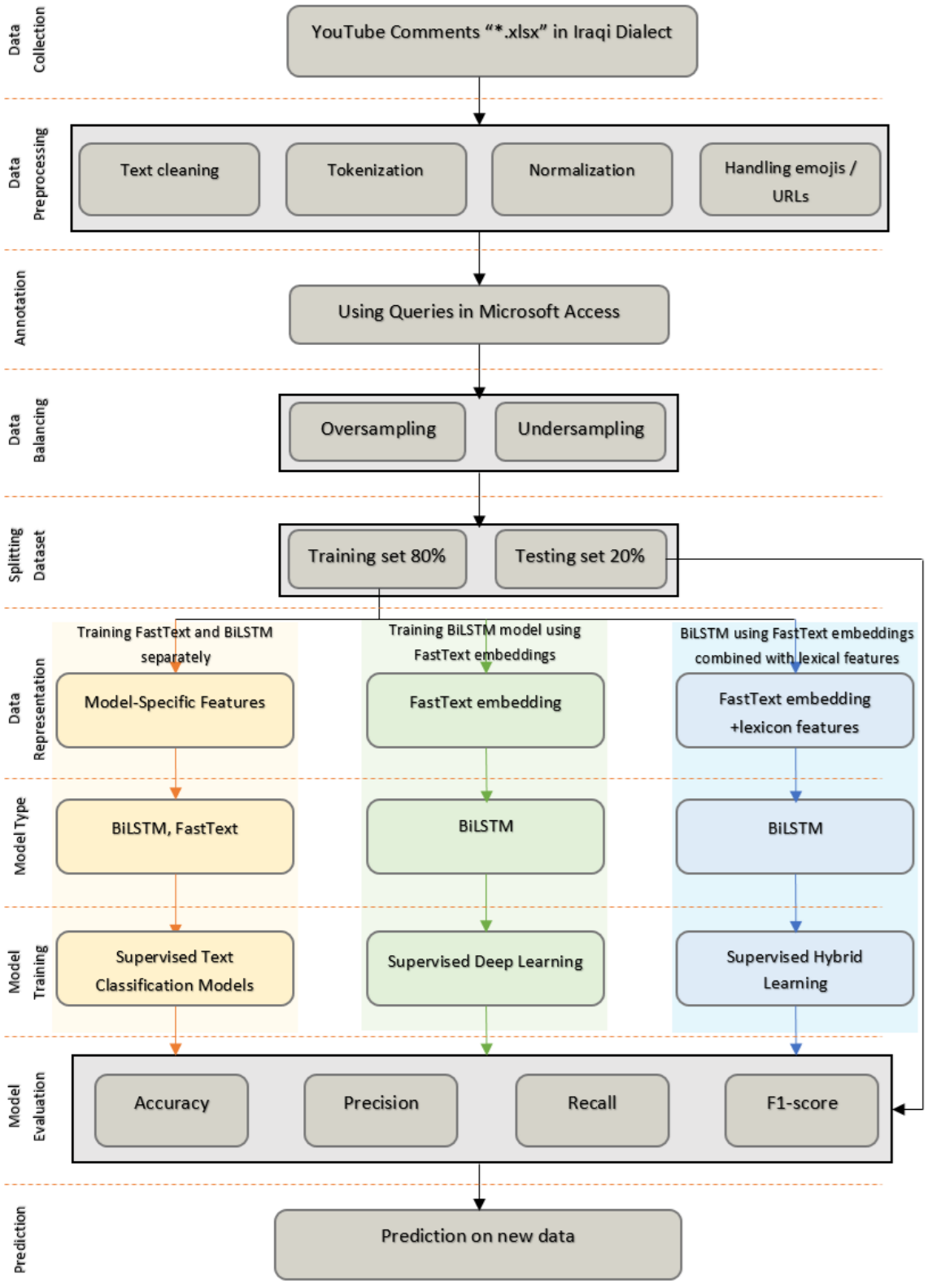

Figure 1. Methodology for detecting hate speech in the Iraqi dialect using three methods**.**

3.5 Training and evaluation

This research applied three distinct ways to train and evaluate two models, namely BiLSTM and FastText, in order to discover which works more successfully, as demonstrated in Figure 2.

1.Training FastText and BiLSTM separately.

Using manually annotated comments in the Iraqi dialect, the machine learning model FastText and the deep learning model BiLSTM were trained as independent classifiers without the use of external lexicons or embeddings. According to the accuracy findings, the deep learning BiLSTM model performed 92.5%, whereas the **FastText** model achieved **96.1%.**

2.Training BiLSTM model using FastText embeddings.

Instead of using the FastText model as a stand-alone classifier, this method trained the deep learning BiLSTM model by using it as an embedding for the same Iraqi dialect dataset. Performance was enhanced by this approach, which produced a **97.3%** accuracy rate.

3.Training BiLSTM model using FastText embeddings combined with lexical features.

This method involved combining FastText embeddings with lexicon-based hate speech characteristics that were sourced from three lexical resources: the first contained offensive words and phrases, the second contained abusive terms, and the third contained expressions that hate speech, which were manually extracted (the offensive word) from the comments themselves. The accuracy of this approach was **92.7%.**

Table 1: Model accuracy results for each strategy implemented.

	Method	**Model**	**Accuracy**
1	Training FastText and BiLSTM separately	BiLSTM	92.5%
		FastText	**96.1%**
2	Training BiLSTM model using FastText embeddings	BiLSTM	97.3%
3	Training BiLSTM model using FastText embeddings combined with lexical features	BiLSTM	92.7%

Accuracy results for each technique are shown in Table 1. By training FastText and BiLSTM independently in the first method, the FastText model outperformed the BiLSTM model, achieving an accuracy of 96.1% and an accuracy of 92.5%, respectively. Therefore, in this method, the FastText model that performs better is taken into consideration for analytical comparison with the following methods. Out of all the strategies, the accuracy of the second approach, which involved training the BiLSTM model using FastText embeddings, was the greatest at 97.3%. However, the third method, which involved training the BiLSTM model with lexical features and FastText embeddings, produced a lower accuracy of 92.7%. Despite the expectation that adding lexical resources would improve performance, it actually caused a little decline. On the surface, this indicates that the second strategy is the most successful, followed by the first and then the third. Complementary evaluation criteria, such training curves and LIME interpretations, however, provide an alternative viewpoint.

- **Training Curve Analysis**

The FastText classifier model, shown in Figure 2, showed a validated accuracy of roughly 96.1%, a fast convergence in the early epochs, and a high training accuracy of over 100%. The flat curve and continuous difference between training and validation performance indicate slight overfitting even if the model shows strong discriminative strength. This could be because of the model's shallow architecture and simplicity.

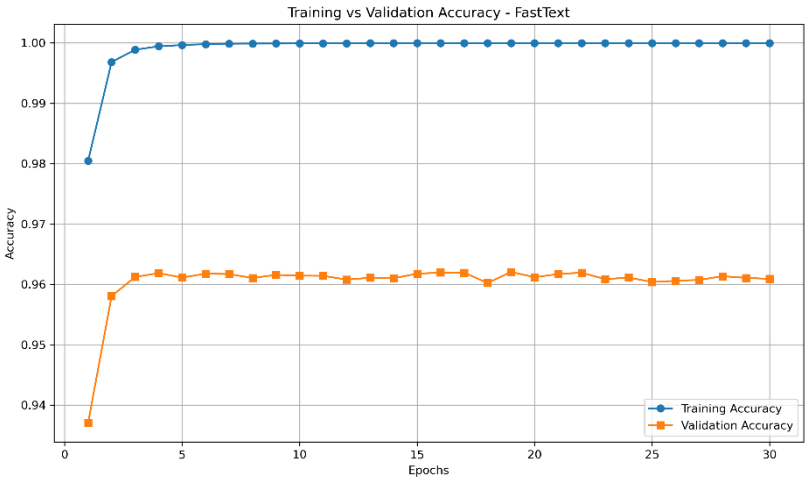

Figure 2. Training Curve of the FastText Model as a Classifier.

Although the accuracy of the BiLSTM model with FastText embedding was the highest in the second technique (97.3%), its training dynamics suggest overfitting.

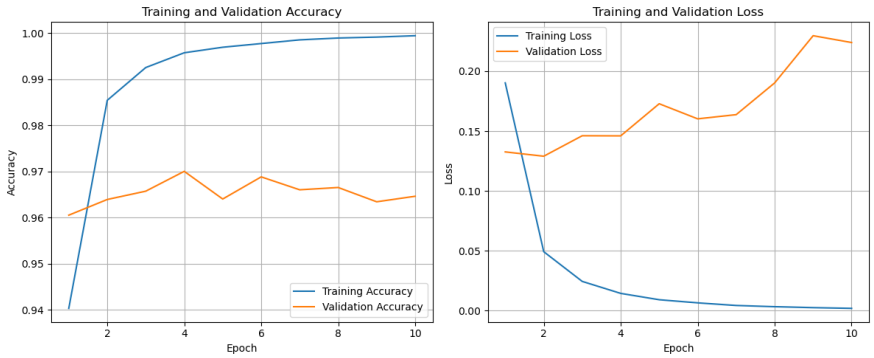

Figure 3. Training Curve of BiLSTM Model with FastText Embedding.

As illustrated in Figure. 3, the training accuracy increases rapidly and nearly saturates at 100%, whereas the validation accuracy plateaus and fluctuates between 96 and 97 percent with an increasing validation loss. The disparity between training and validation performance suggests that the model may have prioritized memorization of training patterns over generalization.

The third method, which combines FastText embedding with domain-specific lexical characteristics (such files containing nasty, abusive, and hateful phrases), exhibits more reliable learning results.

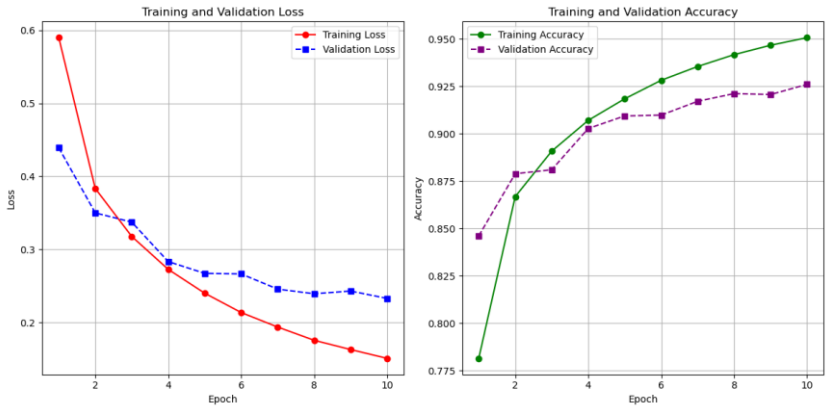

Figure 4. Training Curve of a BiLSTM model using FastText embeddings combined with lexical features.

The accuracy curves in Figure 4 show how training and validation loss decrease smoothly and steadily throughout epochs. In spite of its somewhat lower end accuracy (92.6%) compared to the previous two methods, the third method shows no overfitting and is more robust and generalized. This demonstrates that

while lexical characteristics added complexity, they also improved the model's ability to generalize to new data and helped regularize learning.

- **Interpreter LIME**

The three methods' training curve characteristics are vividly depicted in LIME's visualizations, which also provide insight into the underlying decision-making process of each model.

When taking the sentence " اكطع ادي والراس وافكسها للعين بس لا تكول السنة متحب الحسين من الفلوجة احيكم اخوانا واهلنا بالجنوب وكل العراق " which means "Cut off my hands and my head and slap my eyes, but don't say that the Sunnis do not love Hussein. From Fallujah, I greet you, our brothers and our people in the south and all of Iraq" - This is a metaphor to express the intensity of his love for Hussein. - was evaluated using both the first and second methods, the models classified it as offensive language. However, the actual meaning of the sentence—a normal, natural comment—contradicted this prediction.

Figure 5: LIME Explanation of FastText as Classifier.

Figure 6: LIME Explanation of BiLSTM model with FastText as Embedding.

In both the first method, Independent Multi-Model Training (FastText as a standalone classifier), **Figure 5**, and the second method, Deep Neural Architectures Utilizing FastText Embeddings, **Figure 6**, the LIME visualizations showed a notable dependence on a limited number of highly discriminative words, which were given disproportionately high weights during prediction. One of the best examples is the "افكسها" token, which received a very high attribution score. It is a fairly dependable signal for the model to differentiate offensive speech from neutral or non-offensive speech because it frequently occurs in extremely abusive situations in the training data.

This mistake results from FastText's use of n-gram representations, which break words down into smaller subunits. Segmenting the term "افكسها" revealed that one subunit, was understood to mean "vagina," a clear allusion to a delicate female body part. As

a result, the model classified the comment as **abusive**.

Figure 7. LIME Explanation of BiLSTM model using FastText embeddings combined with lexical features.

The third method, Training BiLSTM model using FastText embeddings combined with lexical features, on the other hand, revealed in LIME a more equal attribution of importance across context terms and lexicon-derived signals, as illustrated in Figure 7. As LIME showed, the highlighted tokens were not just harsh or abusive keywords, but also a mix of lexical clues and contextually relevant terms. The observed balanced training-validation curves, which exhibit minimal overfitting, are in line with this. The algorithm was able to generate context-aware predictions and decrease false positives in neutral or nuanced content by utilizing lexical features.

While the first two methods tended to memorize and overemphasize specific words, the third method combines external lexical knowledge and semantic context to achieve consistent and generalizable performance, which is why LIME interpretations generally support the training curve results.

These findings highlight the need for thorough evaluation procedures in studies on hate speech identification. Deeper insights into model behavior, stability, and decision-making are obtained by combining training curve analysis with LIME-based interpretability in addition to traditional accuracy scores.

4. Conclusion

This research emphasizes how important it is to assess hate speech recognition algorithms using more than just typical accuracy scores, especially when considering the Iraqi Arabic dialect. Although traditional metrics like F1-score, recall, accuracy, and precision are useful benchmarks, they are not enough to fully reflect the range of model behavior. Our results show that depending only on these metrics can result in false conclusions because models may seem very accurate but not be able to generalize well or generate consistent and contextually relevant predictions.

By including training curve analysis, we were able to pinpoint important problems like instability, underfitting, and overfitting during the learning process, providing information on how models change as they are trained. Transparency was further enhanced by LIME-based interpretability, which made the textual patterns and linguistic clues that inform categorization decisions visible. using these supplementary assessment methods, we were able to identify biases, restrictions, and incorrect classifications that were missed using static measures.

The tests, which used FastText and BiLSTM in three distinct methods, further demonstrated that improved accuracy does

not always equate to better performance in practical applications. Rather, it is more dependable to use a balanced evaluation system that incorporates both qualitative interpretability and quantitative performance measurements.

In ultimately, this research highlights how crucial it is to use thorough evaluation techniques in order to create reliable, accurate, and context-aware hate speech detection models.

References

[1] A. Mousa, I. Shahin, A. B. Nassif, and A. Elnagar, "Detection of Arabic offensive language in social media using machine learning models," *Intelligent Systems with Applications*, vol. 22, p. 200376, June 2024, doi: 10.1016/j.iswa.2024.200376.

[2] A. Fonseca *et al.*, "Analyzing hate speech dynamics on Twitter/X: Insights from conversational data and the impact of user interaction patterns," *Heliyon*, vol. 10, no. 11, p. e32246, June 2024, doi: 10.1016/j.heliyon.2024.e32246.

[3] A. Ahmad *et al.*, "Hate speech detection in the Arabic language: corpus design, construction, and evaluation," *Front. Artif. Intell.*, vol. 7, p. 1345445, Feb. 2024, doi: 10.3389/frai.2024.1345445.

[4] R. Ghaly, A. ElKorany, and C. A. Ezzat, "Hate Speech Detection in Arabic Text: Survey," *Procedia Computer Science*, vol. 244, pp. 166–177, 2024, doi: 10.1016/j.procs.2024.10.222.

[5] M. Alkhatib, A. Faisal, F. Alfalasi, K. Shaalan, and A. Mohmed, "Deep Learning Approaches for Detecting Arabic Cyberbullying Social Media," *Procedia Computer Science*, vol. 244, pp. 278–286, 2024, doi: 10.1016/j.procs.2024.10.201.

[6] M. B. Mamun, T. Tsunakawa, M. Nishida, and M. Nishimura, "Hate Speech Detection by Using Rationales for Judging Sarcasm," *Applied Sciences*, vol. 14, no. 11, p. 4898, June 2024, doi: 10.3390/app14114898.

[7] H. Mehta and K. Passi, "Social Media Hate Speech Detection Using Explainable Artificial Intelligence (XAI)," *Algorithms*, vol. 15, no. 8, p. 291, Aug. 2022, doi: 10.3390/a15080291.

[8] K. E. Daouadi, Y. Boualleg, and K. E. Haouaouchi, "Ensemble of pre-trained language models and data augmentation for hate speech detection from Arabic tweets," July 02, 2024, *arXiv*: arXiv:2407.02448. doi: 10.48550/arXiv.2407.02448.

[9] W. Zaghouani and M. R. Biswas, "An Annotated Corpus of Arabic Tweets for Hate Speech Analysis," May 23, 2025, *arXiv*: arXiv:2505.11969. doi: 10.48550/arXiv.2505.11969.

[10] A. Charfi, M. Besghaier, R. Akasheh, A. Atalla, and W. Zaghouani, "Hate speech detection with ADHAR: a multi-dialectal hate speech corpus in Arabic," *Front. Artif. Intell.*, vol. 7, May 2024, doi: 10.3389/frai.2024.1391472.

[11] A. Asiri and M. Saleh, "SOD: A Corpus for Saudi Offensive Language Detection Classification," *Computers*, vol. 13, no. 8, p. 211, Aug. 2024, doi: 10.3390/computers13080211.

[12] Z. A. Abutiheen, E. A. Mohammed, and M. H. Hussein, "Behavior analysis in Arabic social media," *Int J Speech Technol*, vol. 25, no. 3, pp. 659–666, Sept. 2022, doi: 10.1007/s10772-021-09856-6.

[13] R. F. Sabri and N. A. Z. Abdullah, "A Review for Arabic Extremism Detection Using Machine Learning,"

Iraqi Journal of Science, pp. 6617–6630, Nov. 2024, doi: 10.24996/ijs.2024.65.11.35.

[14] W. M. S. Yafooz, "Enhancing Arabic Dialect Detection on Social Media: A Hybrid Model with an Attention Mechanism," *Information*, vol. 15, no. 6, p. 316, May 2024, doi: 10.3390/info15060316.

[15] Z. Tan *et al.*, "Large Language Models for Data Annotation and Synthesis: A Survey," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 930–957. doi: 10.18653/v1/2024.emnlp-main.54

[16] M. Altalhan, A. Algarni, and M. Turki-Hadj Alouane, "Imbalanced Data Problem in Machine Learning: A Review," *IEEE Access*, vol. 13, pp. 13686–13699, 2025, doi: 10.1109/ACCESS.2025.3531662.