Journal of Kerbala University, Vol. 22, Issue 4, December , 2025

IRAQI Academic Scientific Journals

Available at https://www.iasj.net/iasj

Iraqi Academic Scientific Journals

Journal homepage:
https://journals.uokerbala.edu.iq/index.php/UOKJ

Research Article

Machine Learning Framework for Hate Speech Detection in Iraqi Dialect YouTube Comments

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

Article Info
Article history:
Received 19 -11-2025
Received in revised form 2-12-2025
Accepted 15-12-2025
Available online 31 -12 -2025
Keywords: Hate Speech; Iraqi Dialect; YouTube comments; FastText; GNN; RNN; BiLSTM.

Abstrac :

Social media platforms like Facebook, YouTube, and Twitter have witnessed remarkable growth, and the type of data and information shared on these sites has evolved dramatically. Because users of all ages can readily access these platforms, this technological advancement has also been essential in encouraging the spread of hate speech and enhancing its impact on society. Researchers have sought to develop a range of strategies and technology models to detect and mitigate this growing threat.

Even though hate speech identification in English-language literature using Natural Language Processing (NLP) approaches has advanced significantly, research on the Arabic language, especially the Iraqi dialect, is still lacking. This research aims to identify hate speech in the Iraqi dialect by creating a database of more than 150,000 comments taken from YouTube videos about Iraqi topics that have sparked public debate. The gathered remarks were prepared and processed in several steps, including human cleaning. The comments were then divided into four major semantic classes: hate speech, abusive, offensive, and normal.

The efficiency of many machine learning models in processing texts written in the Iraqi dialect was evaluated. Graph neural networks (GNN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Arabic Bidirectional Encoder Representations from Transformers (AraBERT) model, Bidirectional Long Short-Term Memory networks (BiLSTM), and the FastText model were among the models. The outcomes showed that these models performed differently when it came to digesting content in the Iraqi dialect. FastText, on the other hand, recorded a performance rate of 96.1% in both the training phase and in predicting previously unseen remarks, achieving the greatest Accuracy, Precision, Recall, and F1-Score. Therefore, despite its simplicity, the FastText model offers a practical solution for classifying hate speech in different Arabic dialects.

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

Social media has emerged as the most convenient and quick means of international communication. It allows users to participate in various activities, including sharing information, uploading, and commenting. As a result of its growing use in transactions by people, businesses, marketers, and governments, it is a vast source of information [1,2,3].

However, many have misused social media, posting offensive remarks to voice their views or to target a particular group [4]. Some of these individuals used defamation and insults, while others used extremely offensive language. Some of them provoked community hostility, which in some cases developed into intimidation and threats [5]. Their use of pseudonyms instead of their real names contributed to the commission of a number of cybercrimes [6, 7, 8, 9].

Preventing hate speech remains an ongoing challenge and requires systematic and consistent work to reduce its occurrence and minimize its negative effects, with the ultimate goal of removing it from public discourse [5]. Given the ethnic, sectarian, and ideological variety of Iraqi society, the same problem is apparent. remarks on social media sites frequently use language that is hurtful, disrespectful, or encourages hatred.

Hate speech needs to be defined in order to be understood. “Any kind of communication in speech, writing or behavior, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor” [6].

“public incitement to violence or hatred on the basis of certain characteristics, including race, colour, religion, descent and national or ethnic origin” [7] . Another definition is threatening and abusive rhetoric that shows hatred toward a specific group or to another individual, particularly based on their race, color, religion, ethnicity, or even gender [8,9]. Additionally, observe that certain online social media sites (such as Facebook and X platform) have classified hate speech in accordance with their corporate policies [10,11].But as of right now, international human rights legislation lacks a common definition of hate speech. The idea is still up for debate, particularly in light of equality, nondiscrimination, and freedom of speech [12,13,].

Recently, attention has shifted to creating systems that can automatically identify and forecast cybercrimes using artificial intelligence and natural language processing (NLP) [14]. Although there are sophisticated techniques for identifying inappropriate language in English, there is still little study on Arabic [9].There are additional challenges in dealing with the Arabic language related to its writing style. Each country has its own dialect, and most people write in their own dialect, not the official language [14].

In the Iraqi dialect, the differences in accents and vocabulary manipulation are clearly evident, and two approaches can be observed in the comments in the Iraqi dialect:

- There is a wide range of dialects in Iraqi society, with nearly every region having a distinctive speech pattern. For instance, the Arabic word "إذن," which in Classical Arabic means "then" or "so," is pronounced differently in different places. In southern Iraq, it is pronounced "جا," in central Iraq, "لعد," and in other places, "عجل".
- Even among speakers of the same dialect, some people have a tendency to write words according to their pronunciation. For instance, some people write the English phrase "شلونك," which means "How are you?" as "شونك," while others write it as "اشلونك" [1].

Hate speech detection programs have evolved over time, starting from traditional techniques, to the application of different models and algorithms, especially deep learning methods such as transformers and different types of neural networks.

In this paper, we discussed the features of the Iraqi dialect and the difficulties of automated processing, especially when it comes to identifying hate speech. We started by reviewing the features of Iraqi colloquial language and how it differs from Standard Arabic. Then, we described the data collection process, which involved extracting comments from controversial videos about Iraqi issues on YouTube in order to build a local dialect database. We covered the basic steps of natural language processing, including cleaning, normalization, tokenization, and digital representation of text. Next, we trained several machine learning and deep learning models on the data, Evaluating the performance of the models to identify the model that performs best in classifying comments written in the Iraqi dialect. And wrapped up the study by outlining the best model for classifying the data in Iraqi dialect, general conclusions, and this is the main objective of this study.

2. Literature Survey

Recently, research efforts to uncover hate speech in Modern Standard Arabic have increased significantly, focusing on related fields such as sentiment analysis and text classification. Recent research has also begun to focus on Arabic dialects. Unfortunately, there are still very few studies devoted to the Iraqi dialect, leaving a large void in the literature.

A. Detecting Hate Speech in Standard Arabic

By creating a multi-class and multi-label categorization system that divides tweets into five categories of offensive language—bullying, insult, racism, obscenity, and non-offensive— (Mousa et al., 2024) investigated the detection of offensive language in Arabic social media, particularly Twitter. Their method combines deep learning architectures like 1D-CNN and BiLSTM with transformer-based models and conventional machine learning models like RBF and KNN. With an accuracy of 98.4% and an F1-score of 98.4%, the best results were obtained with a cascaded model that started with ArabicBERT and proceeded to BiLSTM and RBF. This study demonstrates how well hybrid models handle the classification of Arabic as an objectionable language [4].

(Alkhatib et al., 2024) used deep learning models, such as CNN, RNN, and hybrid CNN-RNN architectures. About 300,000 tweets made up the dataset, which was first divided into binary categories (cyberbullying vs. non-cyberbullying) before being further divided into six distinct categories of cyberbullying. With an accuracy of 95.59% and an F1-score of 96.73%, the LSTM model outperformed CNN in the binary classification job, while CNN was the best in the six-class classification test with an accuracy of 78.75% [15].

(Bouliche and Rezoug, 2022) introduced a dynamic Graph Neural Network (GNN) model for detecting cyberbullying in Arabic social media, preserving temporal interaction patterns rather than converting data into static graphs [14]. (Daouadi et al., 2024) [16] (Alghamdi et al., 2024) [17] and (Zaghouani and Biswas, 2025) [18]developed a multilabel Arabic tweet corpus for hate speech analysis, AraBERTv2 outperformed other models, and demonstrating strong performance in detecting nuanced hate speech.

B.Uncovering Hate Speech in Arabic Dialects

A corpus of hate speech in Arabic that spans four hate categories and five dialects was produced by (Sharafi et al., 2024) [19]. A collection of hate speech in the Saudi dialect was produced by (Asiri and Saleh, 2024) [9]. The ARABERT model, which has been shown to be successful in identifying hate speech in Arabic dialects, was employed.

Hate Speech Detection in Iraqi Dialect

Behavior analysis on Arabic social media, with a particular focus on the Iraqi dialect, was carried out by researchers such (Abutiheen et al., 2022). They suggested a brand-new categorization technique called the "Identity Classifier" to differentiate between Facebook comments that are wicked and those that aren't. Their research outperformed traditional classifiers tested on the same dataset, reporting an accuracy of 85.4% [1].

after a quick glance at recent research shows a distinct pattern of nations identifying and combating hate speech in user comments published in their own dialect. To increase detection accuracy, this is usually accomplished by using and combining several machine learning and deep learning models. To the best of our knowledge, there are very few studies that focus on identifying hate speech in relation to the Iraqi dialect. Among them is the study conducted by (Abutiheen et al., 2022), which aims to differentiate between wicked and non-wicked remarks on Facebook. The authors' findings were encouraging. The Iraqi dialect is the subject of another study by (Hussein and Lakizadeh, 2025), however this one is restricted to sentiment analysis rather than the identification of hate speech [20]. Consequently, it is imperative to increase studies in order to close this research gap.

3. Characteristics of The Iraqi Dialect

Using Unlike Modern Standard Arabic (MSA), the Iraqi dialect is distinguished by its absence of regular grammatical norms. Even for the same term, there is no single system that governs spelling and pronunciation. Some people follow formal Arabic grammatical structures despite writing in the Iraqi vernacular. For example, they might use the official Arabic plural suffix "شفتوا" to write "شفتو" when conveying the phrase "you saw" in dialect. But instead of writing words according to their formal structure, most people write them phonetically, based on how they are uttered [21]. Following dataset collection and review, these findings can be summed up as follows:

- various spelling variations: Since many words are spelled exactly as they are pronounced, the same statement frequently has various spelling variations. For instance, the insult "طاح حظك" (which means "may your luck fall") is frequently used and can take several forms, such as: "طاح حضك","طاححظك","طاحظك","طاحضك","طايحظك","طايح حضك".
- The letters "ظ" and "ض" are frequently used interchangeably, causing confusion between letters. For example, "محافظ" and "محافض."
- Swapping 'س' for 'ص': Words such as "سيطرة" (checkpoint) could show up as "صيطرة" due to pronunciation-based spelling.
- Spelling errors with plural suffixes: Words that finish in "اذهبوا" (go) can be spelled with or without the final أ, as in "روحوا" and "روحو."
- Final letter variation: Proper names and other nouns that end in "ا,ه,ة,ى" have a variety of spellings. For example, the name "سلمى" can appear as "سلمى," "سلمه," "سلمة," or "سلما."
- Phonological shifts: "ق" frequently takes the place of the letter "ك", as in "يقول" becoming "يكول" or being "ك“ substituted with (ج) as in "كلب" (dog) becoming "جلب".
- Certain sub-dialects alter (ق) to (غ) and vice versa. For example, "ابو الغيرة" becomes "ابو القيرة" and "قسمة ونصيب" becomes "غسمة ونصيب"
- Nunnation simplification: For example, اهلاً وسهلاً is translated as "اهلن وسهلن" when words ending in اً are frequently written with a simple final.
- Prepositions linked with definite nouns: The definite article "ال" is either inconsistently separated or merged with prepositions, producing in forms such as "بلمدرسة", "بل مدرسة", or "ب المدرسة".

- Internal letter rearranging: It's normal to rearrange letters within a word. The word "العلن" (damn) could be written "النعل" instead and the meaning is the same.
- Contractions: abbreviations are commonly used. For example, "شنو سويت" (which means "what did you do?") is sometimes abbreviated to "شسويت."

These linguistic phenomena demonstrate the complexity of written Iraqi dialect, and its diversity poses serious problems for preprocessing and text normalization in machine language processing tasks. Therefore, understanding these complexities is crucial to understanding the challenges of processing and modeling Iraqi dialect data effectively.

4. Methodology

Carefully collecting and annotating data to produce a reliable dataset is essential to accurate and effective model training in hate speech detection. The method that takes the greatest time and effort is this phase. The process of identifying hate speech typically involves several steps. as depicted in Figure 1

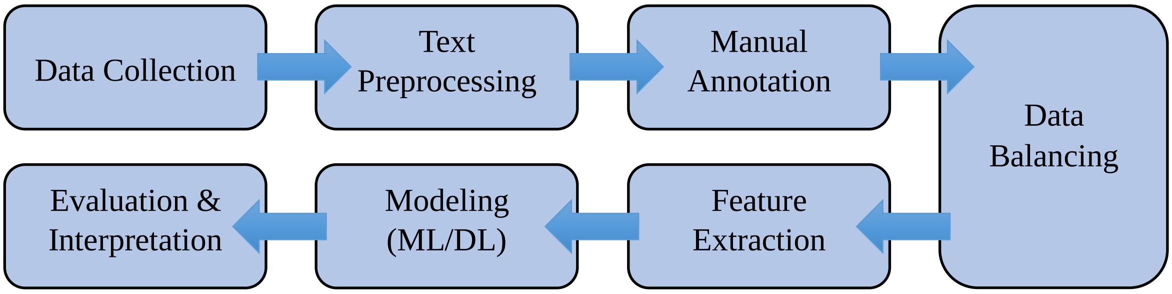

Figure 1: A pipeline for Detect Hate Speech Systems.

A. Data Collection

YouTube was chosen as the main social media platform for this study because of the large number of films that deal with Iraq, which spark discussion and generate many comments in the Iraqi dialect. A single Excel file (.xlsx) was created from 548,661 comments collected from 35 videos. Since they contained a lot of hurtful and non-offensive language, these comments were suitable for building a dataset to detect hate speech. This dataset includes the fields listed below:

● Channel URL ● Name
● Comment ● Time ●Like

Initially The collected comments were contained a significant amount of noise and unlabeled. Many entries were random, duplicated, or included a mixture of various Arabic dialects, English words and letters, and non-informative emojis. This degree of disorganization and irregularity required a very careful and exacting preprocessing step, the raw data had to be cleaned up and made dependable in order to effectively train neural networks and other natural language processing (NLP) models. Table 1. These examples illustrate how noisy and unstructured the original data is.

Table 1: A sample of comments in YouTube in Iraqi dialect

No.	Comment
1	@@gi8iy6vo5gآني عراقي أصيييبل اعذرني ما ارد على الشروكيه
2	@@123vive_algeriaواحد تافه ما عندك مرؤه ولا شرف تف عليك
3	@@btooahmedانت غبي لو مطي ؟ مو ديكلولك القانون يخالف شرع الله المن فاك حلكك بعد ؟
4	@@ci9oيلا انجبي انجبي انتي جايه هنا دكولين كفو وعاشت ايدك وتاكلين تبن وتروحين😂😂😂
5	شوفي صاحب الديكور شكد نغل <<brانتبهو على الوحة خلف المذيع😂😂😂
6	@@خامطهنعالابليس-ص1ب يم فديت الضحكة ي عسل□😚
7	@@fathelalalawi6267ههههههههههههههههههههههههههههههههههههههه شبيككك رحمه لدينك؟
8	-23@@المعيدي حس روحه مقصود بالكلام....طاح حظكم المعدان اللي نصكم نغوله

B. Preprocessing

The Iraqi dialect is known to deviate from the grammatical rules of Modern Standard Arabic (MSA). Consequently, there is currently no model that can accurately processes data in Iraqi dialect. The preprocessing stage required numerous special methods to control the language variation and noise in the collected comments:

1. Removing non-Arabic characters from user comments, including numbers, emojis, English letters, and other symbols that are commonly used.
2. Using Farasa and CAMeL Tools in a Python context, reduce superfluous character repetitions in phrases that roughly correspond to MSA vocabulary (e.g., converting "واااااو هذا يجنننننن" to "واو هذا يجنن").
3. The same tools (Farasa and CAMeL Tools) were used to separate mixed terms whenever possible.
4. Microsoft Access was used to query for and fix words that contained more than three characters, replacing them with the appropriate normalized forms.
5. Usernames listed at the beginning of comments were removed to remove bias and irrelevant stuff.
6. Removing comments written in Arabic dialects other than Iraqi, such as Gulf, Levantine, and Egyptian. To recognize and exclude comments that contained particular words, prior knowledge of these dialects was necessary, like "كده" or "ازيك" (Egyptian dialect) and "لا تحاتي" or "ايش تبي" (Gulf dialect).
7. deleting comments like "🥂🥂🥂🥛🍾🥛🥂" that are made up entirely of lengthy strings of odd emojis.
8. Eliminating utterly meaningless material, such "44خ4خ4خ4خ34خفقخفخففخففخفخفخ3خ34444," which doesn't provide any important information.
9. Eliminated Duplicate comments
10. Character normalization was applied:

- The diacritical symbols being removed.
- The letter hamza (ء) is being removed.
- "ا" was used to refer to all variations of "أ، إ، آ".
- The normalization of "ئ ,ى" to "ي".
- "ؤ" was changed with "و" and "ة" to "ه" [1].

At this stage, a significant amount of effort has been put into data preparation to provide a high-quality dataset that can serve as a reliable basis for natural language processing (NLP) applications, especially in dialect-specific situations such as Iraqi Arabic.

C. Data Classification

A basic task in natural language processing (NLP) is the accurate classification of texts [28], so is important to understand that the offensive terms in the collected comments can be categorized into three main types before beginning the classification process: Hate Speech (instigation to violence), Offensive Language, Abusive Language. To differentiate from the others during the labeling phase, each category needs a clear definition:

- The term "abusive language" describes vulgar or insulting language that is often frowned upon by society and considered inappropriate for use in public settings. These include insults that directly mention sensitive areas of the human body or are sexually graphic.
- Offensive language includes any negative comments or actions intended to ridicule or belittle someone, such as using harsh but non-abusive language or likening someone to an animal.
- Hate speech include expressions that harm individuals or groups by using threats, intimidation, or identity-based targeting. This involves social isolation, inciting violence, and derogatory comments directed toward marginalized groups.

The categorization was determined by the terms' perceived social impact and the extent to which such language is accepted or disapproved of in public conversation and people's reactions and interactions with the information were observed in order to reflect wider social standards. Several comments were found to use several different categories of improper language. In these cases, the comment was grouped based on the most significant violation that was made. For example, a comment was labeled "Abusive" if it used both abusive and insulting language. Since "hate speech" is the most serious category, it was classified as such if it also contained hate speech.

It is also important to recognize that the boundaries between these categories are not always clear and may change in the future. Some expressions that are considered highly offensive may gradually become commonplace, as they are heavily influenced by cultural, temporal, and contextual factors.

Moreover, as was previously mentioned, a large number of words were phonetically transcribed using the writer's own pronunciation, which led to a significant amount of spelling diversity. The term “ لعنة على” (which means “curse upon”), for instance, occurred in a number of irregular forms, including “ لعنهعلا, نعله على, نعلعلى, نعل علا” and others. All of these variants were categorized under a single category rather than being standardized. This method was used to guarantee that the dataset appropriately captures the dialect's linguistic diversity and that NLP models can recognize and handle these expressions independently of their writing style.

Table 2: The comments in Table1 after preprocessing and classification

Comment	Classification
آني عراقي أصيل اعذرني ما ارد على الشروكيه	Hate Speech
واحد تافه ما عندك مروه ولا شرف تف عليك	Abusive
انت غبي لو مطي مو ديكلولك القانون يخالف شرع الله المن فاك حلكك بعد	Offensive
يلا انجبي انتي جايه هنا دكولين كفو وعاشت ايدك وتاكلين تبن وتروحين	Offensive
شوفي صاحب الديكور شكد نغل انتبهو على الوحة خلف المذيع	Abusive
يم فديت الضحكة يعسل	Normal
ههههه شبيك رحمه لدينك	Normal
المعيدي حس روحه مقصود بالكلام طاح حظكم المعدان اللي نصكم نغوله	Hate Speech

D. Data Balancing

After the data was categorized into four groups using Microsoft Access, which made the labeling process quick and easy. The dataset was found to be unbalanced, with roughly 20,000 records listed under the categories of abusive, 30,000 under offensive, 20,000 under hate speech, and 75,000 under normal.

Two methods were investigated to rectify this imbalance:

- By undersampling the majority class (Normal) and somewhat increasing the samples in the other classes, each category was made to include 37,500 records. This resulted in a balanced dataset of 150,000 records overall.
- Oversampling the minority classes to equal the Normal class size, yielding 300,000 records overall, with 75,000 records in each group.

For both dataset versions, a CNN model was used. The model's accuracy on the first (150000) dataset was about 85%, and on the second (300000) version, it was 91%, according to the results. These results showed that several key terms were lost when the Normal class was reduced, which had a detrimental effect on the model's capacity to generalize and make precise predictions about novel inputs. For additional testing across all remaining models, the second strategy—a balanced dataset of 300,000 records—was chosen.

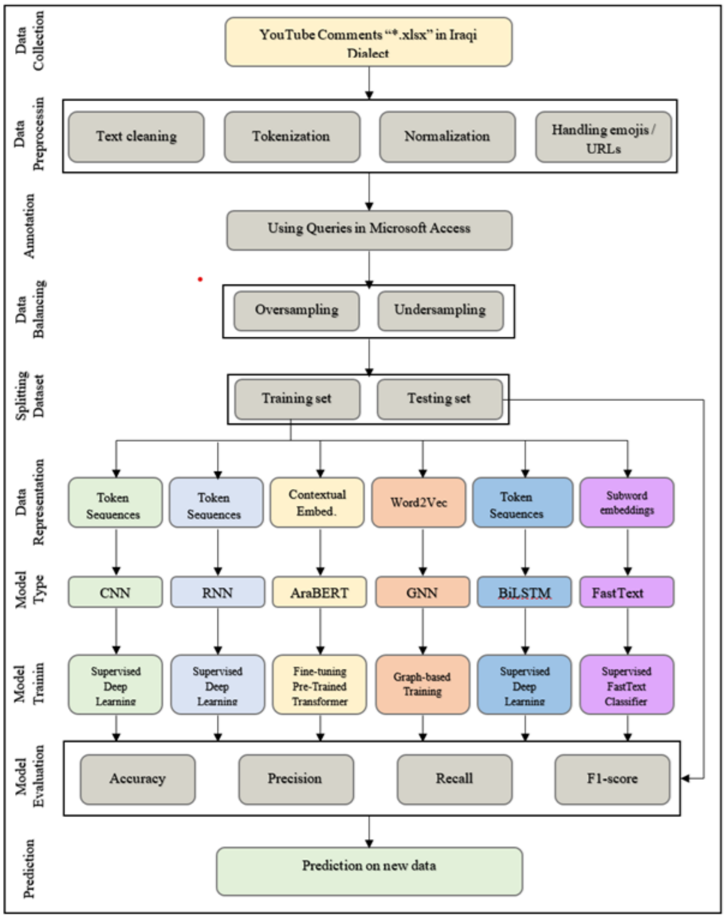

Figure 2: Methodology for Hate Speech Detection in Iraqi Dialect

E. Model Selection

To test their efficacy in managing dialectal textual data and their capacity to categorize hate speech in the Iraqi dialect, a wide range of machine learning and deep learning models were used. As explained below, each model was selected based on its shown capacity to understand noisy or unstructured language:

- **Convolutional Neural Networks (CNNs):** CNNs, which were initially created for image recognition, have demonstrated excellent performance in text classification problems. They are renowned for their computational efficiency and can use convolutional filters to capture local patterns. CNNs successfully recorded brief offensive phrases that are frequently used in dialectal hate speech in this investigation.

- **Recurrent Neural Networks (RNNs):** RNNs are helpful in capturing the temporal connections between words in a phrase because they are made for sequential data. They struggled with lengthy sequences but did rather well with short and medium-length comments. Due to constraints such vanishing gradients and the informal structure of dialectal text, their processing accuracy of the Iraqi dialect was worse than that of other models.

- **Bidirectional Long Short-Term Memory Networks (BiLSTM):** By reading the text both forward and backward and preserving memory over long-term dependencies, BiLSTMs improve on standard RNNs. This bidirectional context helped the model better understand hate expression, resulting in higher accuracy in identifying intricate patterns in Iraqi dialect.

- **Graph Neural Networks (GNNs):** GNNs use nodes in a graph to represent the relationships between tokens. they did not perform well with the Iraqi dialect data. Only 54% accuracy was achieved with the initial training of 30 epochs, and even with the training extended to 250 epochs, the performance only marginally improved to 68%, indicating that structural graph representation for this dataset offers no advantage.
- **AraBERT:** AraBERT, a transformer-based model pretrained on Modern Standard Arabic (MSA), had trouble recognizing the Iraqi dialect's informal and unstructured structure. Its accuracy was less than 60%, indicating that it was not very good at generalizing to dialects that were very different from MSA.
- **FastText:** In the classification of hate speech in the Iraqi dialect, FastText fared better than any other model selected. Because it can manage spelling and phonetic writing differences because to its subword-level embeddings, it is especially well-suited for non-standard material. Additionally, it offered quick training and inference, which improved accuracy and efficiency for noisy, large-scale datasets. In this investigation, FastText had the best categorization accuracy.

F. Training and Evaluation

All models were trained for 30 epochs, with the exception of AraBERT, which was trained for only 3 epochs due to its high computational demands and longer training time. The dataset was divided into 80% for training and 20% for testing to ensure that all classes were proportionately represented in both subsets and the hyperparameters used in training the models (learning rate = Adam default 0.001, batch size = 64, embedding dimension = 128, optimizer = Adam). Table 3 summarizes the classification results for all models under these settings.

Table 3: overview of the training outcomes for the chosen models.

Model	Accuracy	Precision	Recall	F1-Score
RNN	24.9%	6.2%	25.0%	1.0%
GNN	53.6%	53.8%	53.6%	53.6%
AraBERT	55.5%	55.4%	55.5%	55.1%
CNN	90.5%	90.4%	90.5%	90.4%
BiLSTM	92.5%	92.4%	92.5%	92.4%
FastText	96.1%	96.1%	96.1%	96.1%

In Fig. 2 Four important performance metrics—Accuracy, Precision, Recall, and F1-Score—are used to compare several models, including CNN, RNN, BiLSTM, GNN, AraBERT, and FastText.

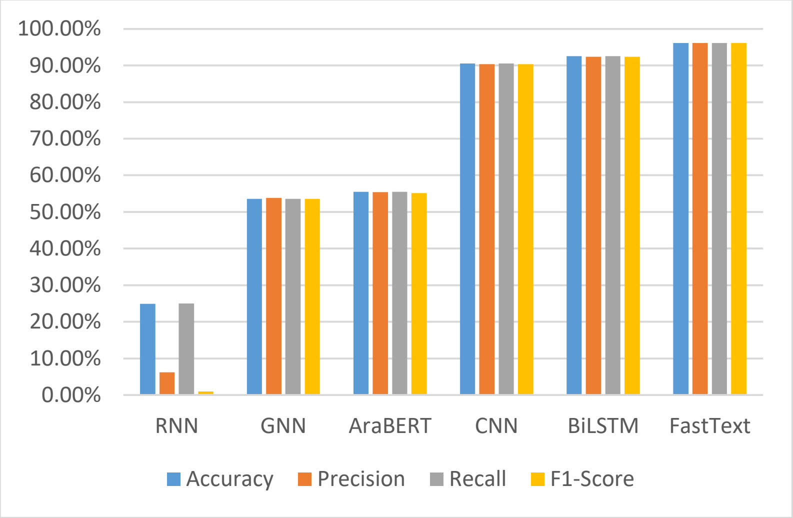

Figure 3: A diagram illustrating the (accuracy, precision, recall, F1) results of the models.

The findings unequivocally show that the FastText model outperformed the others on every metric, proving its effectiveness in dealing with dialectal Iraqi Arabic, which is distinguished by irregular spelling and phonetic variances.

The BiLSTM model ranked second, closely followed by the CNN model. BiLSTM leveraged its bidirectional architecture and long-term memory capabilities, enabling it to better capture contextual relationships within text. While CNN was initially designed for image processing, it has also proven effective at identifying topical patterns.

In contrast, models such as RNN, GNN, and AraBERT showed comparatively lower performance. This suggests limitations in their ability to fully understand the informal and inconsistent nature of Iraqi dialect expressions. Especially the RNN model demonstrated low performance in identifying hate speech in Iraqi Arabic, with a total accuracy of only 24.9%. The incredibly low precision (6.2%) suggests a large false-positive rate and poor forecast reliability. Additionally, the F1-score fell to 1%, demonstrating the model's incapacity to successfully strike a compromise between recall and precision while working with noisy, dialectal text. These findings support the adoption of FastText and BiLSTM as the most effective models for hate speech detection in Iraqi Arabic text, especially when dealing with noisy and morphologically complex data.

5. RESULTS AND ANALYSIS

The FastText model exhibited good classification accuracy across all four categories offensive, normal, hate speech, and abusive—as demonstrated by the confusion matrix in Figure 3. The matrix's dark blue values, which stand for accurate forecasts, are constantly high. For instance, the model showed strong discriminatory capacity by correctly predicting 14,721 out of 15,027 "Abusive" cases and 14,646 out of 14,961 "hate speech" instances.

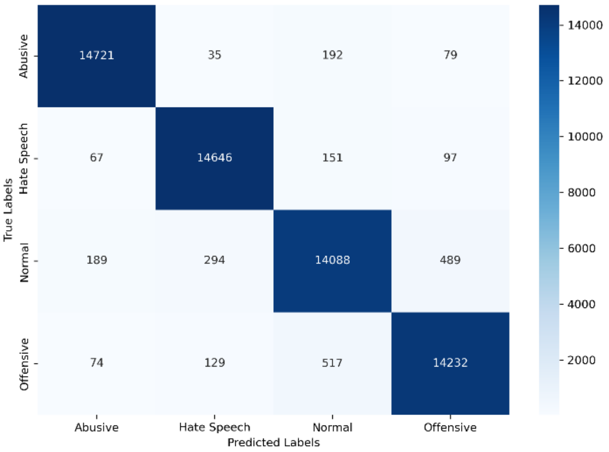

Figure 4: Confusion matrix of FastText model training results.

While Just 192 "Abusive" comments were incorrectly identified as "normal," and 489 "normal" comments were incorrectly classified as "Offensive." However, classification mistakes were small, as seen by the low values outside the dark blue color. This implies that even while the model occasionally mixes up semantically related terms, like "normal" and "Offensive," its overall accuracy and recall are still quite high.

Table 4: Statistical summary

Category	Success rate is approx.	Biggest mistakes with
Hate Speech	97.89%	Normal
Abusive	97.96%	Normal
Offensive	95.14%	Normal
Normal	93.56%	Offensive

The model performed extremely well in the categories of Hate Speech and Abusive. The most difficult was telling between categories "Normal" from "Offensive", as shown in Table 4. Which is to be expected given how similar the context might sound in the Iraqi dialect.

The low total errors show that FastText can effectively manage writing and grammatical changes and comprehend informal dialect in Iraqi dialect hate speech detection tests.

6. LIMITIONS

This research suffers from a number of limitations, including:

1.despite efforts to capture a broad range of lexical and spelling differences, the dataset might not fully represent the variety of written forms utilized in all regions and sub-dialects of Iraqi Arabic.

2.The need for pre-trained models that specialize in dialects, similar to models designed for standard Arabic, and contain as many different written forms as possible for a single word to facilitate its development.

7. CONCLUSION AND FUTURE WORK

In conclusion, identifying and combating hate speech is crucial due to its prevalence and potential to undermine social order. Using natural language processing (NLP), neural network models have shown remarkable effectiveness in identifying hate speech in both Modern Standard Arabic and English. However, their performance tends to suffer when dealing with regional dialects, such Iraqi Arabic, as there are no datasets labeled and no clear orthographic and phonetic norms for the Iraqi dialect. A variety of deep learning and machine learning models were examined after a dataset of YouTube comments in the Iraqi dialect was created, cleaned, and categorized. Notably, the FastText model outperformed the others by achieving an impressive 96% accuracy rate and correctly recognizing the semantic patterns in non-standard text, demonstrating that it is a dependable technique for identifying hate speech in Iraqi dialectal content. In the future, we intend to employ FastText as an embedding layer for deep networks, such as CNN, RNN, etc., rather than as a classifier in order to increase accuracy and gain a deeper understanding of context.

References

[1] 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.

[2] Y. A. Wubet and K.-Y. Lian, "How can we detect news surrounding community safety crisis incidents in the internet? Experiments using attention-based Bi-LSTM models," *International Journal of Information Management Data Insights*, vol. 4, no. 1, p. 100227, Apr. 2024, doi: 10.1016/j.jjimei.2024.100227.

[3] B. Lakzaei, M. Haghir Chehreghani, and A. Bagheri, "Disinformation detection using graph neural networks: a survey," *Artif Intell Rev*, vol. 57, no. 3, p. 52, Feb. 2024, doi: 10.1007/s10462-024-10702-9.

[4] 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.

[5] 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.

[6] U. Nations, "What is hate speech?," United Nations. Accessed: July 08, 2025. [Online]. Available: https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech

[7] "Framework Decision on combating certain forms and expressions of racism and xenophobia by means of criminal law | EUR-Lex." Accessed: July 08, 2025. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=legissum:l33178

[8] S. S. Alaoui, Y. Farhaoui, and B. Aksasse, "Hate Speech Detection Using Text Mining and Machine Learning:," *International Journal of Decision Support System Technology*, vol. 14, no. 1, pp. 1–20, Mar. 2022, doi: 10.4018/IJDSST.286680.

[9] 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.

[10] "Hateful Conduct | Transparency Center." Accessed: July 08, 2025.

[Online]. Available: https://transparency.meta.com/policies/community-standards/hateful-conduct/

[11] "X's policy on hateful conduct | X Help." Accessed: July 08, 2025. [Online]. Available: https://help.x.com/en/rules-and-policies/hateful-conduct-policy

[12] D. A. Das, S. Nandy, R. Saha, S. Das, and D. Saha, "Analysis and Detection of Multilingual Hate Speech Using Transformer Based Deep Learning," Jan. 26, 2024, *Preprints*. doi: 10.36227/techrxiv.170629868.84167256/v1.

[13] M. Chhikara, S. K. Malik, and V. Jain, "Identification of social network automated hate speech using GLTR with BERT and GPT-2 : A novel approach," *JIOS*, vol. 45, no. 2, pp. 315–331, 2024, doi: 10.47974/JIOS-1549.

[14] A. Bouliche, "Detection of cyberbullying in Arabic social media using dynamic graph neural network".

[15] 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.

[16] 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.

[17] 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.

[18] S. Alghamdi, Y. Benkhedda, B. Alharbi, and R. Batista-Navarro, "AraTar: A Corpus to Support the Fine-grained Detection of Hate Speech Targets in the Arabic Language".

[19] 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.

[20] H. H. Hussein and A. Lakizadeh, "A systematic assessment of sentiment analysis models on iraqi dialect-based texts," *Systems and Soft Computing*, vol. 7, p. 200203, Dec. 2025, doi: 10.1016/j.sasc.2025.200203.

[21] F. Husain and O. Uzuner, "Transfer Learning Approach for Arabic Offensive Language Detection System".