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

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

Iraqi Academic Scientific Journals

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

Research Article

YouTube-Based Dataset of User Comments in the Iraqi Dialect

[1],Safaa Hameed Kareem [2], Asia Mahdi Naser alzubaidi

[1,2,]college of computer science and information technology university of kerbala
Kerbala, Iraq

Article Info
Article history:
Received 30 -10-2025
Received in revised form 23-11-2025
Accepted 28-12-2025
Available online 31 - 12 -2025
Keywords: Hate Speech detection, Iraqi dialect, YouTube, natural language processing.

Abstract:

Social media's widespread use has significantly improved communication between people by enabling quick and unrestricted sharing of ideas and opinions. However, some people have also abused this freedom to disseminate hate speech. As a result, abusive language has become more common online, which has a negative impact on society, especially in areas with an elevated level of dialectal variation. In Iraqi society, where almost every province has its own unique dialect that differ greatly from those in other places, a significant challenge to the creation of efficient hate speech detection technologies is this language diversity. A dataset of 548,661 comments in the Iraqi dialect was first gathered from YouTube for this study. The dataset was condensed to contain 120,000 comments after extensive preprocessing, including the removal of random and duplicate comments, etc. The data were then manually classified into four categories: Abusive, Offensive, Hate speech, and Normal language. For investigations involving the identification of hate speech and other associated natural language processing tasks, this augmented dataset provides a fundamental resource. Without a labeled dataset containing actual instances of linguistically offensive material, Natural Language Processing (NLP) algorithms cannot be trained for hate speech identification. Therefore, this dataset was created to train models on Iraqi hate speech.

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 platforms have revolutionized communication, offering users the ability to connect with others across the globe. However, Despite the benefits of these advances, individuals and communities became vulnerable to new forms of harm and verbal aggression that were not common before [1]. So, hate speech on social media is a significant and complex issue [2] that become more prevalent with the rise of digital communication platforms [3],[4].

As a result, identifying hate speech is still difficult. Effectively identifying and analyzing hate speech is challenging due to the variety of dialects and languages, word spelling variations, and inventive sentence structures.

The following are the main challenge to identifying hate speech:

1.Ambiguity in Definition: "Any kind of communication in words, or written actions that attacks or uses derogatory or discriminatory terms with reference to a person or group of people on basis of who they are—in other phrases, based on their religion, nationality, ethnicity, color, racial background, gender, or others identity factors" [5] is the commonly cited term for hate speech provided by the United Nations. There is currently no universally recognized definition of hate speech under the framework of international human rights law [6],[7] , despite the fact that this description might seem all-inclusive. The idea is still up for debate [8], especially when it comes to equality, nondiscrimination, and freedom of speech.

2.Use of Humor: It can be challenging to discern among offensive and non-offensive information when hate speech is included into sarcastic or amusing remarks.

3.Freedom of Expression Issues: The guarantee of freedom of expression may be at stake if the definition of hate speech is broadened.

When working with the Arabic language, these difficulties are exacerbated [9]. Even though it is the official written language, Modern Standard Arabic (MSA) is rarely used in daily conversation. Arabic, the fastest-growing language on the internet, is known for its morpho logical richness, where a single root word can generate hundreds of variations [10]. It is distinguished by rigorous syntactic principles, intricate grammar, and metaphorical language that demands mutual comprehension between the speaker and the listener. Rather, the majority of Arabs use regional dialects, which differ widely in vocabulary and meaning, including Levantine, Egyptian, the Gulf, Yemeni, & Iraqi.

Certain words may be entirely normal in one language but objectionable in another. The word "فشخ" (to hit on the head), for instance, is regarded as neutral in Levantine and Iraqi dialects but highly abusive in Egyptian a language where it is used only as a crude insult. This demonstrates how objectionable language is defined differently depending on culture, which makes it more difficult to identify hate speech consistently.

The difficulty is exacerbated in the case of Iraqi Arabic because of the substantial intra-dialectal variance. The dialects of central and northern Iraq are not the same as those of southern Iraq. In actuality, every province or even tribe may have a distinct lexicon that is solely known within that community. Public Arabic hate speech corpora can be found but rarely concentrate on certain Arabic dialects.

Users frequently utilize indirect or implicit forms of abuse, which makes it hard to determine intent, according to data gathered from social media.

We categorized offensive words into three primary groups based on our findings and manual analysis [2],[11]:

- Abusive language, such as sex words.
- Offensive Wording (such as disparaging or animalistic analogies).

• Hate Speech (e.g., statements that incite hatred or violence against people or groups).

The creation of hate speech detection models specifically suited to the Iraqi dialect was guided by this classification, which also used to organize the annotation process.

Because there are so many films on Iraq that spark debate and generate a lot of comments in the Iraqi dialect, YouTube was chosen as the main social media site for this study. 35 videos yielded a total of 548,661 comments, which were then combined into one Excel file. These remarks were appropriate for creating a hate speech detection dataset because they were full of both offensive and non-offensive language.

Value of The Hate Speech Dataset

Applications in Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning (DL) depend heavily on the utilization of high-quality, well-annotated datasets [12],[13]. In natural language processing (NLP), datasets give models the linguistic context and real-world variability they need to comprehend, produce, and evaluate human discourse. Models are unable to learn syntactic structures, semantic meanings, or pragmatic language use in the absence of representative datasets.

Datasets serve as the basis for both supervised and unsupervised learning tasks in machine learning. They make it possible for models to identify trends, categorize inputs, and extrapolate to previously unseen data. The model's performance in practical applications improves with the diversity and richness of the dataset [14].

Datasets are much more important for deep learning, which uses vast volumes of data to train intricate neural structures. They enable deep models to learn contextual relationships, capture hierarchical language aspects, and adjust to subtle linguistic variances like sarcasm, informal speech, and dialects.

In summary, the accuracy, stock, and durability of NLP and AI systems are directly impacted by the quantity and caliber of datasets [15]. In order to lessen bias and enable linguistic inclusion in technology, dataset building is particularly important for underrepresented languages or dialects. This dataset can be used as a benchmark data set to detect hate speech.

Data Analysis

Upon first examination of the remarks, a number of linguistic and structural issues were identified:

1. It was challenging to understand many, if not most, of the remarks because they were riddled with spelling mistakes and did not follow the guidelines of Modern Standard Arabic (MSA).
2. Some (like "هلتاننال") were written in an entirely arbitrary way.
3. Others were just emoji sequences, like  .
4. A lot of them started with usernames, like "**شخصيتيخطاحمر** اوك براحتج حبيبتي".
5. The Arabic characters "ض" and "ظ" were sometimes used interchangeably (for example, "حض" rather than "حظ").
6. Inconsistency between ص and س, such as using "وصخ" rather than "وسخ" (dirty).
7. Laughter was expressed in a variety of ways, such as " ",خخخخ" ",هههه هههههاااي"," and so on.
8. There were several different types of negative particles, including " ",ما موجود ممموجود" ",ماموجود"," and "مموجود."
9. Words like "now" → "· ,"هسة·" ,"هسا هسى"", and"هس"were characterized by the interchangeability of last letters (ه, ى, ة, and ا).
10. Inconsistent writing was used for particles like "يا" (e.g., " ,"يناس" ,"يا ناس يااناس"").
11. Instead than following normal spelling, grammar, or morphological structures, many users type words as they are phonetically pronounced. For instance: There are several orthographic versions

for the phrase "لعنة على" (curse on), such as *نعل على, نعله على, نعلا على, نعلى على* or *لعن على, لعنه على, نعلهعله*. Likewise, "طايح حظك" (damn your luck) can be found as طايحظك, طيحان حض, طايح حضك, طيححظك, طاحظك and more variations.

A summary Table 1 illustrates key examples of dialectal variation:

Table 1. A sample of different written forms of the same word.

English Mean	MSA	Iraqi Dialect Variants
Now	الآن	هسا – هسة – هسى – هس
Not available	غير موجود	ما موجود – ماموجود – مموجود – موموجود
Please (by God)	بالله	بالله – بللله – بله
Dirty	وسخ	وسخ – وصخ
O people	يا ناس	يا ناس – ياناس – يناس
Laughter	ضحك	ههه – خخخخ – هاااا – هههاايي – هههعع
At the university	بالجامعة	بالجامعة – بلجامعة – بل – جامعة
Except	الا	إلا – اله – الة
For him	لهُ	له – اله – إلا – الة

Additional observations include:

- Some dialects replace "غ" with "ق" and vice versa; for example, "قرفة" means "room" instead of "غرفة," and "غاتل" means "murder" instead of "قاتل."
- "ك" is occasionally used in place of the letter "ق", also "ج" instead of "ك", creating ambiguity that can only be cleared out by understanding the entire sentence context. For example, " شلونك يا كلب" may not signify "dog" but rather "قلب" (heart).
- The ending "ا" is inconsistently used following plural verbs (for example, "شفتوا" versus "شفتو").

In general, Iraqi Arabic writing is phonetic, controlled more by pronunciation than by formal grammar. Dropped letters and asymmetrical word structures result from this. Some people even try to purposefully obfuscate information in order to avoid detection, like:

1. Dividing words into their constituent letters such as "ا ن ج ب و ل ك" that mean “shut up”.
2. Writing words that are only partially complete but nonetheless have meaning.
3. Completing the word with pictures or emojis.
4. Adding symbols, numerals, or English characters to express meaning.
5. Making derogatory references to people or social situations without using direct language.
6. Including unpleasant language in a lighthearted environment, which makes it difficult to categorize purpose.

These trends show how difficult it is to identify hate speech in Iraqi Arabic and emphasize the need for models that can deal with noise, dialectal diversity, and colloquial language.

Hate speech dataset description

Most previous research work has utilized Twitter, due to the short length of its posts and data availability [1], while this work collected YouTube comments—despite requiring more preprocessing—because it is not investigated in any arranged manner in the past studies about Iraqi dialect.

Additionally, YouTube comments are long and can contain many different types of hate speech in one comment, making them more valuable for analysis.
The creation of a dataset involved gathering user feedback from 35 publicly accessible YouTube videos that addressed contentious issues pertaining to Iraqi society. These films were specifically chosen because of their high levels of engagement and the fact that they contain both offensive and non-offensive statements in the dialect of Iraq. 548,661 comments in all were collected and saved in Excel (.xlsx) format.

The dataset includes the following fields:

- Channel URL
- Name
- Comment
- Time
- Like

Table 2. A sample of comments in YouTube in Iraqi dialect.

No.	Comments
1	اااف اني طويلة بس موحيل😥😥😥😥 لعد راح اعنس كلها تحب😂😂😂😂😣😣😣😥😥😥القصيرة
2	😂😂انجب ابو الوصخ طبعا هذا الصدك تضوجون منه @@iopezkamal4869
3	امجنات-غ5ش شنو معدان كلبي واذا تلبس بنطرون اهم شي جوه العبايه شويه ثقفو نفسكم@@
4	ءٍ“ ما تطيريد /ثكلن بناتي شويه ثكلن /(برابيگ)💜ʦخـہاٰبـہہٰ ☻ - الدره طولها متر
5	تفشلين اخر من يتكلم انت، انت مو عراقيه و لاتعرفين شي لا تمضرطين و الي لازم ينطرد 🤣🤣🤣😭 المذيع الي ما عنده اخلاق
6	😒💔بيلا-ـث6غ ششهآلألففاظظ عـآبــتت ههل ششككلل، @@
7	عبالك مزوجة حرامي االبنيه المتزوجة من تجي زعلانه امها تگلها جبتي ذهباتج وياج صدك عود ليش😁👌👌ه 😂😂
8	البصرهجوهرةالخليج نصيحة مني روح نام آخر زمن يجي شروكي ما يعرف راسه من رجله يكوم @@ يتفلسف 😂

It is noteworthy, therefore, that no preliminary classification of whether a comment is offensive or not was rendered. The majority of the comments are written in Iraqi Arabic, and they are distinguished by the following:

- a high frequency of emojis, symbols, and numerical characters.
- frequent use of Arabic and Latin characters.
- usage of other regional dialects (such as Egyptian or Levantine).
- a large number of spelling errors.
- high redundancy and repetition among the comments.

Before any preprocessing procedures are applied, Table 2 shows a sample of raw, unprocessed comments to demonstrate the linguistic variety and noise within the dataset.

Data Preprocessing

It is well known the Iraqi language does not follow Modern Standard Arabic's (MSA) grammatical standards. As a result, no model is currently available that can analyze data in Iraqi dialect with any degree of accuracy, not even at a 50% threshold. In order to manage the linguistic variety and noise in the gathered

comments, the preprocessing stage necessitated many unique steps:

1.Elimination of non-Arabic characters, such as emojis, English letters, numbers, and other symbols frequently seen in user comments.

2.Reducing unnecessary character repeats in terms that approximate MSA vocabulary by using Farasa and CAMeL Tools in python environment (for example, changing "وااااااو هذا يجنننننن" to " واو هذا يجنن").

3.Where feasible, combined words were separated using the same instruments.

4.Words with more than three characters per word were queried and corrected using Microsoft Access, which replaced them with the proper normalized forms.

5.To get rid of bias and unrelated content, usernames that were listed at the start of comments were eliminated.

6.Comments in other Arabic dialects, such as Gulf, Levantine, or Egyptian, were not included. Prior understanding of Arabic dialect terms like "جذي" "كويس," "ازيك," and "يلبألك" was necessary for this. Some examples of comments in Arabic dialects: “ الحياة صعبة أوي على أي حد عايز يفضل إنسان استخفرالله ياربي ايش هذه ياربي والله “ ,”كويس مايستاهل العراگ جذي يصير بي ❤❤❤”.

7.To maintain objectivity and prevent model bias, comments containing offensive remarks about explicitly named political or religious figures have been deleted.

8.We found and removed duplicate comments.

9.If isolated characters were unnecessary or did not change the semantics of the comment, they were either eliminated completely or combined with nearby words to complete their meaning.

10. Normalization of characters was used:

- Removing the diacritical markings.
- All forms of "أ، إ، آ" were united as "ا".
- “ى, ئ” was normalized to "ي".
- "ؤ" was unified with "و" in instances of synonymous usage.
- Removing the letter hamza (ء).
- "ة" was changed to "ه" [16].

11. Where feasible, spelling adjustments were made and consecutively repeated terms were eliminated.

12. In order to address dialect-specific irregularities and guarantee high-quality data, manual inspection and correction remained essential even with the employment of all these automated technologies.

To guarantee the precision and dependability of the outcomes, a considerable amount of work was spent on data preparation at this point. In order to generate a high-quality dataset that may be a reliable basis for natural language processing (NLP) tasks—especially in dialect-specific contexts like Iraqi Arabic—this meticulous procedure was necessary.

Data Classification

Before starting the classification process, it is crucial to recognize that the objectionable terms found in the gathered comments can be divided into three primary groups:

1.**Abusive Language**,

2.**Offensive Language**, and

3.**Hate Speech (incitement to violence)**.

During the annotation stage, each category must have a precise definition in order to be distinguished from the others:

- **Abusive Language** refers to offensive or vulgar language that is typically disapproved of by society and deemed unsuitable for public conversation. These include insults that are sexually explicit or make direct references to delicate parts of the human body.
- **Offensive Language** involves any remarks or derogatory behavior meant to make fun of or denigrate someone, such as making animal analogies or employing harsh yet but not abusive.
- **Hate Speech** includes expressions that use intimidation, threats, or identity-based targeting to hurt people or groups. This includes insulting remarks made to underprivileged groups, incitement for violence, or social isolation.

Even in international legal situations, as was previously said, there is no widely agreed-upon definition of hate speech. This begs the question: How were the study's comments categorized? The classification was determined based on the perceived social impact of terms or vocabulary, and their acceptance or rejection in public discourse. This was observed through people's reactions and interactions with the information, reflecting broader social norms. By searching for offensive vocabulary used in comments, the comment was classified based on whether or not it contained that offensive word.

It was discovered that some comments used multiple inappropriate language types. The comment was categorized in these situations according to the most serious infraction that was committed. A comment was classified as "Abusive" if it contained both offensive and abusive language, for instance. It was categorized as "hate speech" if it also contained hate speech, since this is the most serious category.

It's also important to recognize that the lines separating these groups are not always clear and may change in the future. While certain expressions may change in meaning or intensity, others that are today deemed extremely unpleasant may progressively become more commonplace. As a result, the concept of offensive language is dynamic and greatly impacted by contextual, temporal, and cultural elements.

Microsoft Access was used for the data categorization process, and update queries were created to change the classification field according to whether or not particular terms or phrases appeared in each comment. Comments that contained inappropriate language were flagged as such (e.g., hate speech, insulting, or obscene). On the other hand, remarks devoid of any offensive material were classified as "normal".

Furthermore, as was previously mentioned, the classification took into consideration every spelling variation of important terms, such as "لعنة على" and " طاح حظك," which might occur in a variety of orthographic forms. A well-structured and dialect-specific dataset was produced by this thorough process, which makes it appropriate for testing and training natural language processing (NLP) models on Iraqi Arabic.

Table 3 displays the comments listed in Table 2 after preprocessing. Following data normalization and cleaning, the comments were divided into four groups according to their semantic cues and linguistic content, which reflected the distribution of terms among the groups: normal , hate speech, abusive , and offensive speech.

Table 3. The comments after preprocessing and annotation.

Comment	Classification
اف اني طويله بس موحيل لعد راح اعنس كلها تحب القصيره	Normal
انجب ابو الوصخ طبعا هذا الصدك تضوجون منه	Offensive
شنو معدان كلبي واذا تلبس بنطرون اهم شي جوه العبايه شويه ثقفو نفسكم	Hate Speech
الدره طولها متر خايبه ما تطيرين تكلن بناتي شويه تكلن برابيك	Abusive
تفشلين اخر من يتكلم انت مو عراقيه ولاتعرفين شي لا تمضرطين والي لازم ينطرد المذيع الي ما عنده اخلاق	Abusive
شهالالفاظ عابت هل شكل	Offensive
البنيه المتزوجه من تجي زعلانه امها تكلها جبتي ذهباتج وياج عبالك مزوجه حرامي صدك عود ليش	Normal
نصيحه مني روح نام اخر زمن يجي شروكي ما يعرف راسه من رجله يكوم يتفلسف	Hate Speech

Data Availability

The dataset utilized in this research is openly accessible on Kaggle. It consists of YouTube user-generated comments in the Iraqi Arabic dialect with hand classification labels. Three main columns are included in the Excel-formatted dataset:

- ID: each record's unique sequence identification.
- Comment: The YouTube comment's unformatted content.
- Classification: the label that corresponds to whether the comment is normal, hate speech, abusive, or offensive.

The dataset, which includes a significant amount of annotated social media content, is designed to aid Arabic natural language processing (NLP) research, particularly for tasks like dialectal analysis and hate speech identification.

The following URL allows you to see and download it without charge:
https://www.kaggle.com/datasets/safaahameed1978/iraqi-dialect-youtube-comments-for-hate-speech

Conclusion

Because the Iraqi dialect deviates from Modern Standard Arabic (MSA) and lacks established grammar or spelling rules, it poses a substantial linguistic and technological difficulty for hate speech detection. It is particularly challenging for conventional NLP techniques and models to attain dependable performance in Iraqi Arabic due to its informal, phonetic, and very diversified nature, as well as users' inventive language manipulation.

The dataset underwent a thorough, multi-phase preprocessing procedure to solve this, combining a great deal of manual inspection with automated methods using programs like Farasa, CAMeL Tools, and MS Access. Character normalization, dialectal inconsistency elimination, spelling correction, emojis and symbols filtration, and dialect-specific word variant management were the goals of these procedures. Particular attention was also paid to removing text in non-Iraqi dialects, noise, and derogatory comments to specific people.

Even with the advanced Arabic natural language processing (NLP) methods available today, human involvement is still necessary to capture the complex and wildly fluctuating Iraqi dialect. The cleaned, normalized, and annotated dataset that results offers a strong starting point for further research into dialectal Arabic classification of texts and hate speech detection.

The significance of dialect-aware processing and the demand for customized options in low-resource, dialect-rich dialects like Iraqi Arabic are emphasized in this work.

References

1. Ahmad, A. *et al.* Hate speech detection in the Arabic language: corpus design, construction, and evaluation. *Front. Artif. Intell.* **7**, 1345445 (2024).
2. Mamun, M. B., Tsunakawa, T., Nishida, M. & Nishimura, M. Hate Speech Detection by Using Rationales for Judging Sarcasm. *Applied Sciences* **14**, 4898 (2024).
3. Fonseca, A. *et al.* Analyzing hate speech dynamics on Twitter/X: Insights from conversational data and the impact of user interaction patterns. *Heliyon* **10**, e32246 (2024).
4. Gongane, V. U., Munot, M. V. & Anuse, A. D. Detection and moderation of detrimental content on social media platforms: current status and future directions. *Soc. Netw. Anal. Min.* **12**, 129 (2022).
5. Nations, U. What is hate speech? *United Nations* https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech.
6. Ghaly, R., ElKorany, A. & Ezzat, C. A. Hate Speech Detection in Arabic Text: Survey. *Procedia Computer Science* **244**, 166–177 (2024).
7. Jahan, M. S. & Oussalah, M. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing* **546**, 126232 (2023).
8. Chhikara, M., Malik, S. K. & Jain, V. Identification of social network automated hate speech using GLTR with BERT and GPT-2 : A novel approach. *JIOS* **45**, 315–331 (2024).
9. Mousa, A., Shahin, I., Nassif, A. B. & Elnagar, A. Detection of Arabic offensive language in social media using machine learning models. *Intelligent Systems with Applications* **22**, 200376 (2024).
10. Asiri, A. & Saleh, M. SOD: A Corpus for Saudi Offensive Language Detection Classification. *Computers* **13**, 211 (2024).
11. Hashmi, E. & Yayilgan, S. Y. Multi-class hate speech detection in the Norwegian language using FAST-RNN and multilingual fine-tuned transformers. *Complex Intell. Syst.* **10**, 4535–4556 (2024).
12. Zeng, M. *et al.* Data Quality Enhancement on the Basis of Diversity with Large Language Models for Text Classification: Uncovered, Difficult, and Noisy. Preprint at https://doi.org/10.48550/arXiv.2412.06575 (2024).
13. Wang, Z. *et al.* Diversity-oriented Data Augmentation with Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2502.11671 (2025).
14. Nguyen, D. & Ploeger, E. We Need to Measure Data Diversity in NLP -- Better and Broader. Preprint at https://doi.org/10.48550/arXiv.2505.20264 (2025).
15. Pluščec, D. & Šnajder, J. Data Augmentation for Neural NLP. Preprint at https://doi.org/10.48550/arXiv.2302.11412 (2023).
16. Abutiheen, Z. A., Mohammed, E. A. & Hussein, M. H. Behavior analysis in Arabic social media. *Int J Speech Technol* **25**, 659–666 (2022).