Journal of Techniques, ISSN: 2708-8383, Vol. 7, No. 3, 2025, Pages 53-63
DOI: https://doi.org/10.51173/jt.v7i3.2695

JOURNAL OF TECHNIQUES

Journal homepage: http://journal.mtu.edu.iq

RESEARCH ARTICLE - ENGINEERING (MISCELLANEOUS)

IQAD: Iraqi Arabic Dialect Dataset for Multi-Regional Dialect Classification Using Conventional and Machine Learning Approaches

Noora Aljubouri[1], Naderi Hassan[1*]

[1]Department of Software, Computer Engineering College, Iran University of Science and Technology (IUST), Tehran, Islamic Republic of Iran

[*] Corresponding author E-mail: naderi@iust.ac.ir

Article Info.

Article history:

Received
22 May 2025

Accepted
17 August 2025

Publishing
30 September 2025

Abstract

The work's main contribution is creating a dataset for specifying Iraqi Arabic dialects from written texts. With the increase of Iraqi dialectal Arabic usage across social media platforms, accurate dialect identification has become an important step for such tasks as sentiment analysis, social media monitoring, and linguistic studies. We collected, annotated, and prepared normal text data: 53,146 unique text samples taken from social media, divided into three major dialects in Iraq: Middle, Western, and Southern. The lexical variability of the corpus is 78,582 unique tokens. The dataset was passed through preprocessing to clean and prepare it for classification-based tasks. To verify the quality of this dataset, we carried out experiments with two approaches for the classification: a dictionary-based methodology and a TF-IDF-based SVM classification. The SVM outperformed the dictionary-based classifier by achieving 74% accuracy and F1-score, whereas the classifier peaked at 63.6% accuracy and 63.4% F1 score. The results show the effectiveness of the dataset in supporting dialect classification tasks and its potential for use in future Iraqi Arabic NLP applications and research.

Publisher: Middle Technical University

Keywords: Machine Learning; Dialect; Language; Classification; SVM.

1. Introduction

The increased usage of regional dialects on digital platforms has raised interest in more accurate systems of dialect identification, especially for the Arabic language. Arabic dialects, from one region to another, pose several challenges to many NLP tasks, such as machine translation, sentiment analysis, or speech recognition. However, Iraqi Arabic is unusually rich in phonological, lexical, and syntactic variations across regions [1]. Yet much research interest has been placed on Arabic NLP, but the Iraqi dialect has limited focus. Most existing tools and datasets center around Modern Standard Arabic or other dialects with more attention, such as Egyptian or Levantine Arabic. This lack of structured data and resources becomes a major issue in developing machine-learning models related to the Iraqi dialect [2].

The motivation behind this study is to provide a high-quality dataset for the written dialects of Iraqi Arabic. Through collecting and preprocessing data from real-world sources such as social media, it is hoped that this will give way to further work on dialect classification and dialect-aware applications.

In order to test the usefulness of this dataset, we used it to perform dialect classification by two methods: one rule-based approach (dictionary method) and machine learning using term frequency-inverse document frequency (TF-IDF) features and support vector machine (SVM). Comparative results validate the usefulness of the dataset while opening a new door toward machine learning approaches for Iraqi dialect classification tasks [3].

2. Literature Review

Significant efforts have been focused on classifying linguistic accents or dialects in recent years. Modern Standard Arabic (MSA) has traditionally been the focus of most Arabic natural language processing (NLP) research efforts. Nonetheless, there has been an increased curiosity about Dialectal Arabic (DA), focusing on dialects like Egyptian or Lebanese. Several studies were conducted, producing valuable results that can be summed up as follows:

Tibi & Messaoud (2025) proposed an adaptive deep learning model for Arabic dialect identification. A novel Multi-Scale Product Analysis (MPA) was employed for feature extraction, and a Hamilton Neural Network (HNN) was used for classification. The model showed better results when compared to other approaches. However, the system struggled with dialects with similar phonetic similarities (e.g., Moroccan vs. Algerian, Iraqi vs. Libyan), and the dataset's limited speaker variability did not permit generalization. Therefore, future work is proposed to enhance the feature representations further and treat the related performance trade-offs with non-causal models [4].

Nomenclature & Symbols			
ML	Machine Learning	TF-IDF	Term Frequency-Inverse Document Frequency
MSA	Modern Standard Arabic	NLP	Natural Language Processing
DA	Dialectal Arabic	SVM	Support Vector Machine

Yassir Matrane (2023) discussed sentiment analysis on dialectical Arabic (DA). It acknowledged that grammatical, vocabulary, and syntactical variations across dialects constitute the biggest hindrance in polarity classification. Another finding of the work was that it highlights the important steps that affect machine learning models of dialect sentiment analysis, ranging from text annotation to preprocessing, feature extraction, and approaches adopted. Furthermore, the study presented challenges and open issues in Arabic dialect sentiment analysis research [5].

Alsarsour and Mohamed (2018) presented a new large, manually-annotated multi-dialect dataset of Arabic tweets that is publicly available. The Dialectal Arabic Tweets (DART) dataset contains some 25K tweets annotated via crowdsourcing. It is well-balanced over the five main groups of Arabic dialects, namely Egyptian, Maghrebi, Levantine, Gulf, and Iraqi. The paper presents the pipeline of constructing the dataset-from crawling tweets, which match a list of dialect phrases, to having the tweets annotated by a crowd. The dataset's quality was evaluated from two angles: inter-annotator agreement and accuracy of the final labels. Results revealed that such measures were substantially higher for the Egyptian, Gulf, and Levantine dialect groups but lower for Iraqi and Maghrebi dialects, indicating the difficulty of manually identifying those two dialects and, consequently, automatically identifying them [6].

According to Keleg and Walid (2023) introduced the Arabic Level of Dialectness (ALDi) as a continuous measure and released the large-scale AOC-ALDi dataset with 127k+ annotated sentences. They used a BERT regression model to quantify dialectness. Key issues include Arabic diglossia, dialectal variation, and moderate inter-annotator agreement. The study highlights gaps in dialect diversity, gender bias, and generalizability, paving the way for richer Arabic dialect modeling [7].

Alansari (2023) developed a deep learning model for detecting and classifying Standard Arabic. The idea was to create a model combining CNN and RNN to catch the semantic, syntactic features. They divided their approach into six stages: NLP, feature engineering, neural networks, language models, optimization, and evaluation. The work tried to beat the limitations of the old methods. Verifying the model's accuracy and capability is explicitly stated as future work. This highlights the ongoing gap in fully realized, tested AI systems for the complex and varied Arabic dialect [8].

Sadat et al. (2014) showed the application of Naïve Bayes classifiers and the character n-gram language model by analyzing which models work best in various social media contexts. With an accuracy of 98%, the classifier the authors trained using the character bigram model could distinguish between the 18 distinct Arabic dialects [9].

Adnan and Emran (2021) developed a manually annotated Arabic sentiment corpus from social media and tested five classifiers, with SVM performing the best (F1-score: 83%). The paper solved the unavailability of Arabic sentiment datasets in the open domain and showed an effective way of pre-processing and extracting features. Still, the researchers note the challenges of dialectal diversity, small-sized corpora, and richer linguistic features. While their study fills a significant gap in Arabic NLP, it also pinpoints the need for more dialect-aware, scalable tools for sentiment analysis [10].

2.1. Trends in Arabic language processing

Recent trends in Arabic language processing are pointing toward a gradual shift from the exclusive study of Modern Standard Arabic to more fine-grained studies of Dialectal Arabic, given that dialects are increasingly being used in real-life communication [5-7]. To accommodate this transition into multi-dialectal study, researchers have developed large, annotated multi-dialect corpora such as the DART and AOC-ALDi datasets [6, 7]. To grasp semantic and syntactic complexities in Arabic, deep learning techniques are used nowadays, including CNN-RNN architectures and, more recently, transformer models such as BERT [4-8]. Continuous measures of dialect, such as ALDi, were introduced to move from the regular practice of traditional binary classification. However, many challenges remain, like dialectal overlaps, diglossia challenges, low inter-annotator agreement, and dataset biases such as gender imbalance. This shows a necessity for broader dialect coverage and better feature engineering in future work.

The most often used supervised algorithms in the literature are K-nearest neighbors (KNN), Naïve Bayes (NB), decision trees, and support vector machines (SVM). This method was applied to the Arabic language in several attempts. For instance, MSA sentence-level subjectivity and sentiment analysis using the SVM on a small annotated corpus gathered from news stories was described in [11].

2.2. Challenges in Iraqi Arabic text recognition

Identifying Iraqi Arabic in text form presents challenges because it differs from Modern Standard Arabic (MSA) and other dialects in its sound, word structure, and vocabulary features. The complexity of Arabic text recognition systems stressed the need to consider dialect differences to ensure accurate transcription. Besides, the variations in phonetics in the different regions of Iraq make it harder to create reliable text recognition models.

Iraqi Arabic has significant sound differences from modern Arabic and other regional dialects. These differences include changes in tone, stress patterns, and vowel and consonant sounds. For example, Iraqi Arabic pronounces some consonants differently from MSA. Additionally, individual speakers and dialect factors can also affect vowel pronunciation due to the geographical influence. To classify Iraqi Arabic text, it's crucial to keep these variations in mind [12].

Due to its unique properties, various difficulties have to be considered. Sentence and word structure modifications, including plural construction, different word spelling, and verb conjugations, are known as morphological variants. Moreover, the lexical variants describe the linguistic and terminological distinctions between Iraqi Arabic and standard modern Arabic.

Each Iraqi dialect has morphological, lexical, and phonetic characteristics that differ from the others. The differences are due to numerous elements, including geographic location, religion, ethnicity, social level, and historical influences. These variances are not usually considered

in the Arabic dataset; therefore, machine learning models trained on public datasets find distinguishing dialect language difficult. What complicates the detection procedure is, for example, when the speakers switch between MSA, dialectal varieties, and other languages, a practice known as code-switching. These challenges are identifying and considering these morphological and lexical variations when accurately transcribing printed Iraqi Arabic text [12].

3. Theoretical Background

3.1. Machine Learning (ML)

Artificial intelligence (AI), which includes machine learning, allows computers to automatically learn from data and get better at it without needing to be explicitly designed. ML algorithms are used to identify trends in data and forecast future outcomes [13].

The area of computer science known as machine learning uses prior experience to gain information and apply that knowledge to future decision-making. Machine learning is at the nexus of statistics, engineering, and computer science. Machine learning aims to generate an unknown rule from instances or generalize a discernible pattern. Although machine learning may be roughly divided into three types, these categories can be blended depending on the circumstances to provide the intended outcomes for specific applications. Those types are unsupervised, supervised, and reinforcement learning [14]. For this work, supervised learning will be used to classify the dialect samples into the target dialects.

3.2. Supervised Learning

Supervised learning has been considered a fundamental training paradigm in machine learning, where data is provided in a labelled form so that the trained model can predict an outcome corresponding to known labels. Supervised learning works best for text classification tasks, such as Arabic dialect identification and sentiment analysis, which are core to this study. Each input sample in supervised learning is attached to a target label, either categorical in classification or continuous in regression [15].

This learning strategy fits our work because Arabic NLP training data sets (e.g., comments labelled for dialect or posts tagged for sentiment) already contain human-annotated labels. These labels become the guidance for the learning algorithm to identify linguistic patterns in the data, after which it produces accurate results for previously unseen data. Moreover, in recent years, some top contender models trained through supervised learning approaches, accompanied by state-of-the-art Arabic tasks such as SVMs or classifiers built upon BERT architectures, have yielded tremendous results [15]. Fig. 1 shows the flow chart of the supervised learning process.

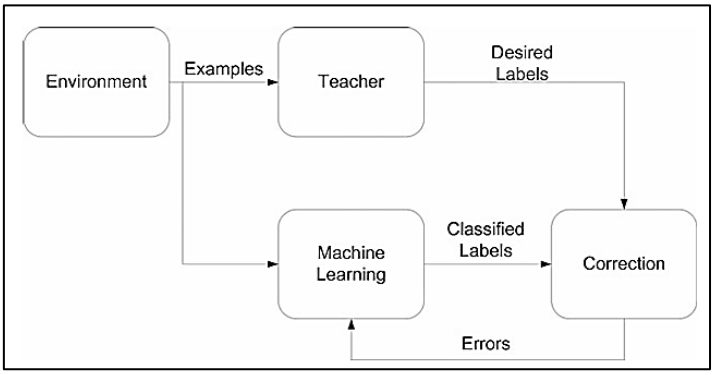

Fig. 1. The supervised learning process [15]

3.3. Support Vector Machines (SVM)

A supervised learning algorithm with strong power and performance has been used for classification and regression. The introduction of support vector machines (SVMs) was by Vapnik and Cortes in the mid-1990s. SVM techniques have become immensely popular because of their efficiency in dealing with high-dimensional data and their strong theoretical underpinnings [16].

At its heart, SVM attempts to find the best hyperplane to separate data points of different classes with the utmost possible margin. Margin is the distance between the hyperplane and the nearest data points from both sides, which are called support vectors. The basic idea is that a larger margin leads to better generalization and sturdiness against unseen data [16].

SVMs have been used intensively in many text categorization tasks like spam detection and sentiment analysis. They are significant to text as they can classify very high-dimensional sparse data [17].

3.4. Evaluation metrics

Several widely used evaluation metrics for classification tasks, such as accuracy, precision, recall, and F1-score, are used to assess the performance of different machine learning and deep learning models. These metrics show how the models perform when evaluated using several criteria and their accuracy in classifying the Iraqi Arabic dialects from text input, for example. These parameters are significant concerning the overall correctness of the classification on a balanced data set and how it deals with unbalanced or multilabel dialect classes situation that commonly occurs in Arabic NLP [18].

- Accuracy: is the percentage of the total number of test instances that are correctly classified. It is a good parameter to be used when one has a balanced data set. But its reliability drops in cases where some dialects seize a larger portion of the corpus [5].
- Recall: or sensitivity measures the model's ability to correctly identify all members of a specific class, e.g., detecting Iraqi dialect utterances. This becomes very important in our context since some classes may be under-represented to the extent that they might not be considered. As shown in equation (1) [13].

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Postives} + \text{False Negatives}} \quad (1)$$

- Precision: refers to how correctly instances for a dialect can be predicted. It becomes most important when the boundaries between dialects become fuzzy and carry the grievance of high misclassification cost, as shown in equation (2) [13].

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Postives} + \text{False positives}} \quad (2)$$

- F1-Score: a harmonic mean between precision and recall, balances between them when dealing with class imbalance. This imbalance is common in dialect datasets, where some dialects have more data than others. The F1 score is determined by both precision and recall, as shown in equation (3) [13].

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (3)$$

Among these, this work particularly uses the accuracy and the F1-score, which allows us to observe from a more pertinent perspective the model's capacity to classify instances from both minority and majority classes. These metrics form a tighter set to evaluate model performance in Arabic dialect classification.

3.5. Text Encoding using Term Frequency-Inverse Document Frequency (TF-IDF)

Turning raw data into numerical vectors as structured data is known as data encoding. A numerical vector is always provided as the input to use machine learning algorithms for real-world situations. To apply machine learning algorithms, the raw data must be converted into numerical vectors because it is nearly always provided in a different format.

The TF-IDF method was chosen to encode text in this study. It has effectively converted unstructured text into the numerical vectors needed by machine learning models, including SVM. Out of the many encoding methods applicable to natural language texts, from BoW and word embedding techniques to one-hot encodings, TF-IDF is very appropriate for capturing the relevance of terms within and across documents, resonating with the concept of dialect classification. The TF-IDF consists of two parts:

- The Term Frequency (TF) shows how often a word pops up in a document. It is found by dividing the number of times a word appears in a document by the total number of words. The idea behind TF is that the more a word shows up in a document, the more it matters to that document [19].
- The Inverse Document Frequency (IDF) measures a term's importance across a corpus. It is calculated as the logarithm of the ratio of total documents in the corpus to documents containing the term. The IDF component down-weights terms that appear across many documents, as they lack discrimination power [19].

Previous research successfully used this method to represent Arabic text for classification tasks, like when paired with a machine learning model like SVM. The studies by Adnan and Emran in 2021 demonstrated high accuracy using SVM on Arabic dialect and sentiment datasets on classification tasks. This is due to TF-IDF's ability to emphasize subtle lexical differences between dialects [10].

TF-IDF is simpler, easier to understand, and has exhibited outstanding success over time. It is one of the most well-established and explainable approaches to this problem. Combined with SVM, which can handle highly sparse data well in high-dimensional contexts, it provides a fine and proven baseline for Arabic dialect classification [19].

4. Methodology

The study follows a structured methodology consisting of five steps: (1) data collection from Iraq-region-based Facebook pages by using an automated approach supplemented by a manual method; (2) data cleaning and preprocessing to filter out noise, such as names, emoji, or non-Arabic characters; (3) feature extraction and encoding of the features through a dictionary-based method and the TF-IDF vectorization method; (4) training and selection of classifiers while evaluating the performance of a dictionary-based classifier and a straightforward linear SVM; and (5) evaluation of the selected models concerning classification metrics such as accuracy and F1-score. These steps address particular problems of dialectal variation and very high lexical overlap among regions to make for a reliable Iraqi-dialect classification from social text.

4.1. Data collection and preprocessing

The data collection process was conducted over 4 months, during which user-generated content was gathered from Facebook. It was chosen as the primary data source due to its widespread usage among Iraqi dialect speakers and its suitability for capturing informal, region-specific language. Comments were collected from users residing in three major regions of Iraq: the south, the middle, and the west. This was verified by inspecting publicly shared user information and post content.

4.1.1. Web scraping methods

Two main methods were employed for data acquisition:

- Chromium comment scraper extension: A browser-based extension that automatically extracts Facebook comments from a given internet link of the target post (see Fig. 2). It can save these comments in various formats: CSV, JSON, or Excel (XLSX). This method was used for posts with high user engagement, giving rise to a large corpus of usable comments.
- Manual extraction: The process moved towards manual collection in posts with low comment counts or unsupported by automated tools. The page's content was saved into text files, and then custom Python scripts were used to extract Arabic text while eliminating non-Arabic content using regular expressions and other preprocessing.

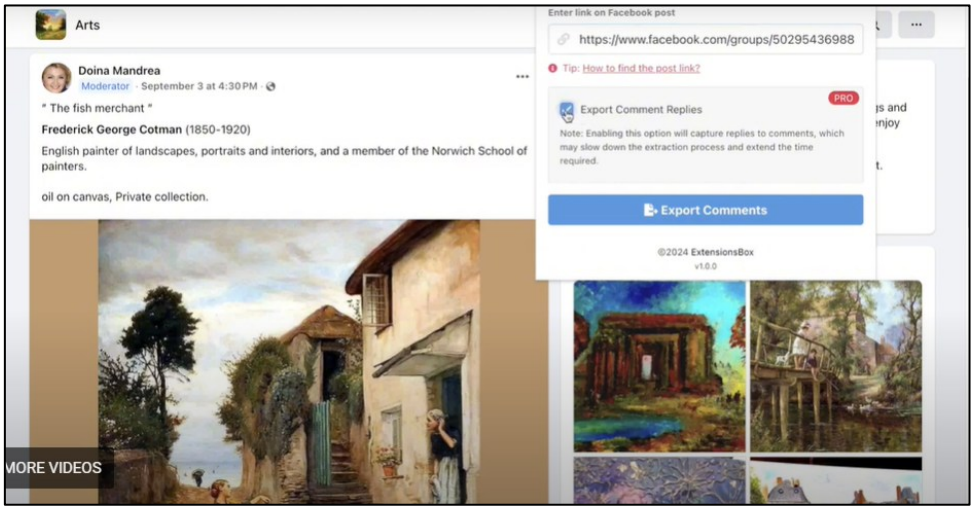

Fig. 2. Chromium comment scraper extension, which works by providing a link to the target page, and scrapes the comments into either JSON or CSV format

4.1.2. Identifying dataset sources

The dataset is designed to collect dialectal comments from users in Iraq's southern, middle, and western regions. These three regions were chosen due to the prominent linguistic differences perceived in each dialect and their active participation in regional Facebook discussions.

To ensure regional restrictions and to avoid nearly useless noise, the data was not collected from mainstream content or other national-level pages such as the one for main news or politics, as these pages invariably attract users from all over the country. On the contrary, pages dedicated to local communities that cover narrowly focused events at the neighborhood level, like neighborhood news or small-town announcements, were targeted. The content was inspected to ensure that most samples meet the requirements.

Cross-regional contamination was kept under high supervision. One region's users often comment on posts from another due to reasons like internal migration or just plain interest. Therefore, before the dataset is finalized, every candidate page was manually reviewed, and in some cases, user profiles and comments were also sampled.

All automatic scraping tools like APIFY were deliberately avoided because they do not offer fine-grained filtering and might have resulted in the extraction of unqualified data or data of mixed-region origin. The Chromium scraper and manual selection were the only methods applied when extracting the comments, further ensuring the dataset's high degree of regional purity.

4.1.3. Sample cleaning and preparation

The cleaning and preparation were done through multiple phases using Python scripts to automate the processing, which included the following:

- Removing non-Arabic characters using the "re" package: The script used a built-in package called "re", which performs a regex pattern search to remove, for example, specific characters.
- Removing common Iraqi names: A list of familiar Iraqi person and title names was developed and extended upon observation to include nicknames, which are common where users avoid writing their true names. Each comment was inspected, and names were removed. This is important since names do not contribute to the contextual meaning of the text sample.
- Removing numbers and punctuation: Some comments contained some numbers and punctuation, which did not contribute to the meaning of the text and were thus decided to be removed. Besides, numbers and punctuation are common to all dialects, even outside the Iraqi dialect.
- Removing HTML links and URL: Some comments contained web links from different sources for advertisement purposes, for example. The links were removed since they always consisted mainly of non-Arabic characters.
- Removing emojis: Facebook comments tend to contain Emojis often to express emotions. Those Emojis are not region-specific; the world commonly uses them, and thus were removed, since they do not contribute to the dialect classification.
- Removing repeated characters like in ("تمااااام") example, which does not have any grammatical or language structure, or a specific length. Some texts contain repeated characters which used to modify the expression, like when ("هلووووو") is written to express excitement in welcoming someone. Those repeated characters were removed to obtain the original word.

The cleaned comments were then combined into a text file for each region while keeping the original raw, uncleaned data for future needs. The text files are then labeled for the specific region.

An example of the collected raw samples (Fig. 3) demonstrates the number of unwanted characters that needed to be cleaned and treated:

Fig. 3. Raw text before processing as a word cloud

4.1.4. Dataset analysis

53,146 comments were collected and distributed across Iraq's three main dialect regions: the south, middle, and west. After normalization and filtering, the combined dataset contained a vocabulary size of 78,582 unique words. The distribution of comments per region is illustrated in Fig. 4, where the western region has the largest number of comments, followed by the southern and then the middle region. Therefore, considering the three classes, the dataset produced a somewhat balanced split, with a few samples from the western zone perhaps skewing towards a larger number of comments.

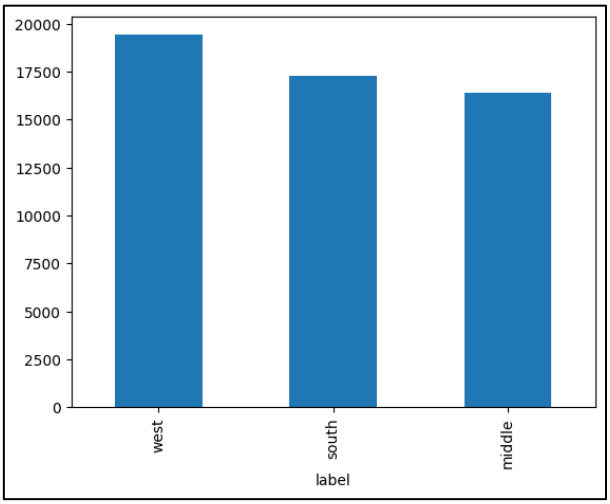

Fig. 4. The number of comments for each region

Since the length of comments varies widely, we considered the lengths of such entries. The dataset comprises single-word entries with comments of up to 195 words. The overall distribution of lengths is shown in Fig. 5, displaying a skewed distribution with an accumulation of shorter comments and a very long tail of long ones. Non-uniformity should be addressed in the preprocessing pipeline stage, particularly during tokenization and model training, to avoid biases towards shorter samples.

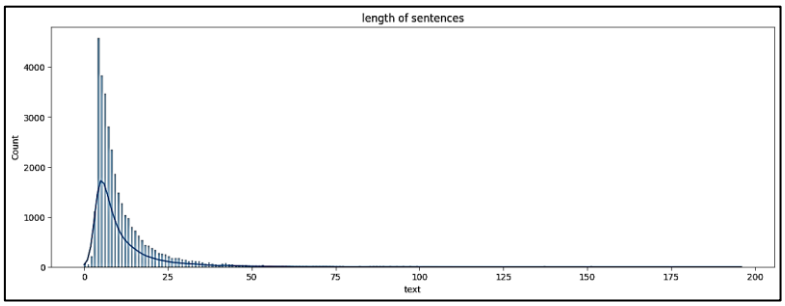

Fig. 5. The length distribution of comments

Going from larger to smaller regional views, one sees the very same non-uniform length patterns. The plots shown in Figs. 6 to 8 are length distributions for the south, west, and middle regions. Shorter comments dominate the datasets in all three regions, having larger comments that are relatively fewer in number. This differential in comment length may affect the classification performance if not normalized adequately.

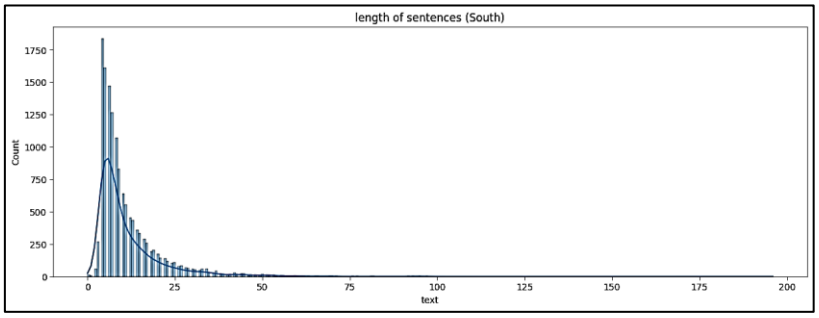

Fig. 6. The length distribution of comments from the south region

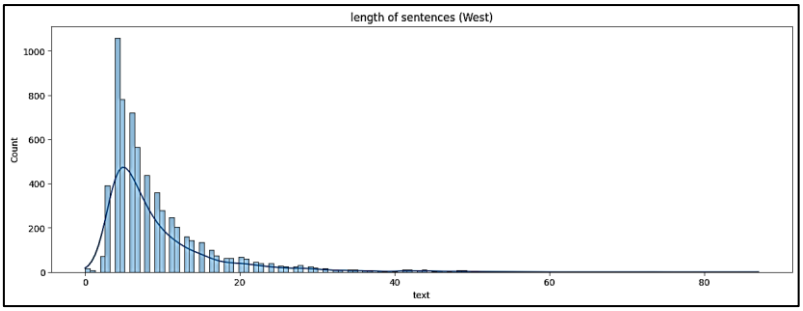

Fig. 7. The length distribution of comments from the west region

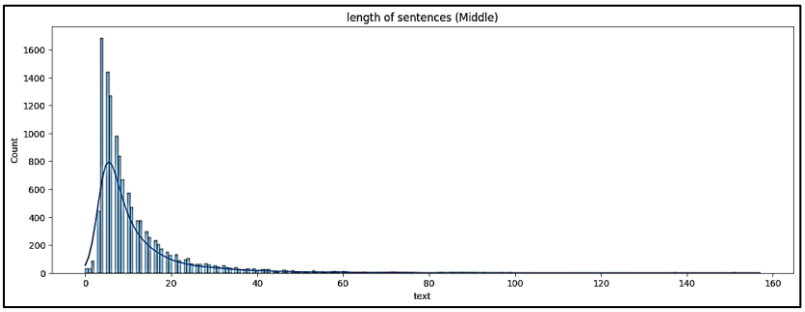

Fig. 8. The length distribution of comments from the middle region

In addition to comment length, it performed vocabulary overlap analysis across the dialect regions to understand lexical similarities and potential challenges in classification. The following statistics summarize the total and unique word counts per region:

- Southern dialect: 37,785 words (28,121 unique)
- Western dialect: 38,831 words (29,167 unique)
- Middle dialect: 38,720 words (29,056 unique)

Here, it observes significant lexical overlap among the three dialects, notwithstanding the regional differences, as 70.4% of the words are in common among all three regions. It implies that the vocabulary is very shared, making it difficult to classify the dialects, especially when written text is considered, since the prosodic and phonetic cues, such as pronunciation or vocal accent, are removed.

Percentage-wise, regarding common words used within each region:

- In the Southern dialect, 70.9% were common.
- In the Western dialects, 70.5% were common.
- In the Middle dialects, 69.7% were common.

This seemingly exhaustive use of the shared vocabulary presents huge challenges when differentiating dialects just from word use. Hence, the classification model will depend largely on syntactic patterns, word usage context, and rare regional expressions, which amount to about 29.6 percent of the entire dataset. These findings reveal the challenges tied to the task and the great potential for extracting more refined linguistic features.

4.1.5. Comparison with existing Arabic dialect datasets

While the Dialectal Arabic Tweets dataset and AOC-ALDi have positively impacted the development of Arabic dialect studies, the dataset under consideration fills a unique gap. Meanwhile, DART considers tweets from five primary dialect areas (Egyptian, Gulf, Levantine, Maghrebi, and Iraqi), treating the Iraqi dialect as a single class without regard to intra-Iraqi regional varieties. This is where our dataset comes in as the first to classify three Iraqi dialects-southern, middle, and western- from regionally validated samples. Moreover, the AOC-ALDi dataset views dialectness as a continuous linguistic variable comprising considerable amounts of MSA. In contrast, our focus is entirely on dialectal, informal written Arabic, gathered from public regional Facebook pages, paving the way for context-specific and informal usage examples. Our

data are labeled for classification so supervised learning can be applied directly, unlike ALDi, where dialect intensity is scored on a regression basis. Table 1 shows a comparison between previous datasets and ours.

Table 1. Comparison between datasets from previous works and our dataset

Dataset	Dialects Covered	Focus	Granularity	Source	Iraqi Coverage
DART	5 major (incl. Iraqi as a single group)	Dialect classification	Tweet-level	Twitter	Treated as one group
AOC-ALDi	MSA + Dialects	Dialectness regression	Sentence	Online comments	Not regionalized
our Dataset	3 Iraqi dialects (South, Middle, West)	Dialect classification	Comment	Facebook (regional)	Region-specific

4.2. Model selection

Two overall different modeling approaches were considered for the dataset evaluation-the first being based on rules, or a non-machine learning method, versus a machine-learning approach. The ultimate aim of this comparison is not simply to gauge classification performance but to evaluate how well the dataset aids dialect identification under two opposing paradigms.

The reasons for having these two approaches evaluated are:

- A rule-based (conventional) model uses hand-crafted rules, keyword dictionaries, and statistical frequency-based features. They are often employed in low-resource settings where labeled data is very scarce, and they afford interpretability and simplicity. By running such a model, we get a baseline performance measure and can determine whether dialectal patterns within the data can be directly captured through lexical or syntactic cues.
- Machine learning models learn patterns from the data to capture explicit and implicit data features; they outperform their rule-based counterparts if given sufficient labeled data and may generalize to examples never encountered before. Testing present-day ML models on this dataset provides insight into which data can support data-driven generalization and abstraction. SVM was selected due to its effectiveness when dealing with high-dimensional sparse feature spaces such as those produced by TF-IDF. It is computationally efficient and is considered well enough for small-scale settings. While alternatives such as logistic regression or Naïve Bayes could also be used, the SVM has ranked above both in prior Arabic dialect classification efforts [10].

4.2.1. Dictionary-based classifier model

In this concept, a dictionary for each region was created, where the unique words found from the comments from a target region were collected and saved. Besides, another Dictionary contained the common words seen in all the languages.

Model concept:

- Comments are tokenized and then passed to a function that collects the score for each word or token.
- Words that belong to a specific region got a high score, and depending on the number of words that belonged to a specific group, a total score was given.
- Each sentence will have three different scores (middle_score, west_score, south_score), which then the maximum score is then searched and corresponds to the model prediction.

Fig. 9 shows a Dictionary-based model, where a dictionary of the unique words for each dialect is created. Each new text will be checked against those dictionaries to see how many words exist in the dictionaries, and the one that shows the most word count that belongs to the dictionary is the winning class and the correct prediction.

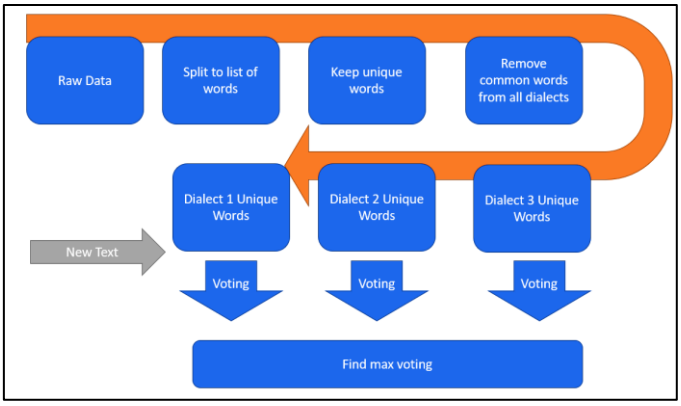

Fig. 9. The concept of the Dictionary-based model, where a dictionary of the unique words for each dialect is created

4.2.2. TF-IDF SVM

A linear SVM model for classification was combined with a TF-IDF vectorizer, which generated numerical vectors based on the statistical measures of the words and their frequency in the document, where the frequency values of the words in the vector represented the feature. The input vectors were then provided to the SVM model, which contained a linear kernel with better performance when dealing with sequential data. The SVM maps the feature vectors into another hyperplane and works to maximize the margin between the three classification classes' hyperplanes. The linear SVM was used from the Sci-Kit Learn package and fitted on a preprocessed dataset. The SVM model was wrapped with a one vs rest classifier (OneVsRestClassifier from Sci-Kit Learn [20]) to enable the model to do multiclass classification. The model's performance was evaluated using a classification report tool from the same package, which showed the metrics for the model and each class.

4.3. Evaluation

The selection of Accuracy and F1-score is commonly proposed to assess the effectiveness of the trained models and the dataset's quality. These metrics ensure a balanced view of the model's performance, primarily since many dialectal datasets reflect imbalanced or overlapping classes.

Accuracy indicates the percentage of samples correctly classified regarding the simplicity of assessing general performance. However, the actual usefulness of accuracy may become insignificant when the classes are non-perfectly balanced or share a high number of overlapping features, as is usually encountered in dialect classification.

F1-score combines precision and recall to provide a single number measuring the utility of classification models on datasets wherein the boundary among classes (dialects) is subtle or where one class may dominate the data distribution.

- Validation process

The evaluation was conducted post-training. Both conventional and machine learning techniques received the dataset prepared after cleaning, preprocessing, then splitting into training and validation sets. Preprocessing involved normalization, tokenization, and vocabulary filtering. The same split was carried through onto both systems to ensure the evaluations were done in comparison with each other.

The entire modeling and evaluation pipeline is depicted in Fig. 10, from raw data to model evaluation. The flowchart depicts each step, beginning with raw data and cleaning, passing through preprocessing via training, toward evaluation with accuracy and F1 score.

This evaluation framework thus ensures that these results hold reproducibility and interpretability among models, whether rule-based or machine learning. The details of these observed performance metrics, along with challenges of dialect overlap, vocabulary similarities, and migration effects impacting regions as far as the observed figures, are discussed further in the results section.

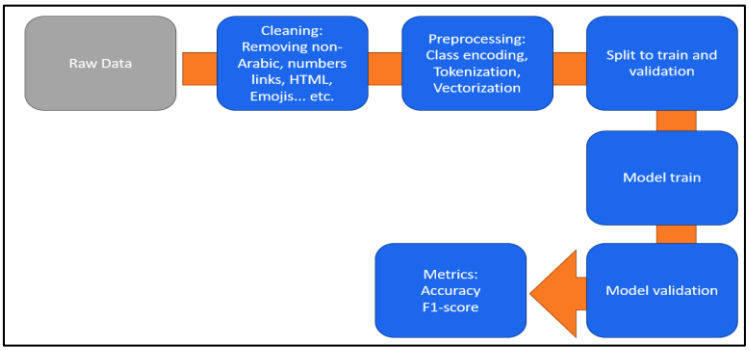

Fig. 10. The flow graph of the model concept, starting with raw data, cleaning, preprocessing, and ending in evaluating the model using Accuracy and F1-score metrics

5. Results and Discussion

This section presents a comparative study on the two dialect classification models and their relative potentials in classifying the discussed Iraqi Arabic dialects. Furthermore, it addresses the limitations of the study and possible future works. In the concluding part, the main thrust of contributions made by this work concerning related works in the field is highlighted.

5.1. Model performances and comparisons

Two types of models were implemented to emphasize the potential of the dataset and to investigate the dataset structure:

- Dictionary-Based Classifier (Non-Machine Learning Approach)
- TF-IDF SVM Classifier (Machine Learning Approach)

Both models were trained on 80% training data and tested on 20% test data to maintain fairness during the comparison. The data were split similarly so that each model saw the same training and validation sets.

5.1.1. Dictionary-based classifier

The dictionary-based model uses a list of specified words for each dialect. During training, it constructs a dictionary of regionally distinguishing vocabulary. During classification, it infers the labels based on the frequency of words in the input compared to each dictionary in the regions. The model achieved an accuracy of 63.6% and an F1-score of 63.4%.

The strength of this model is presented in its simplicity, with the low demand on computational resources. However, this model cannot generalize beyond its vocabulary or account for linguistic nuances, especially in cases where context or sentence structure becomes imperative. It cannot deal with out-of-vocabulary (OOV) words, which can be detrimental in any real-world, evolving language setting like social media.

5.1.2. TF-IDF + SVM classifier

The machine learning model transformed the text data into numerical features using TF-IDF vectorization, which were then fed into a linear SVM classifier. The model attained 74% accuracy and 74% F1-score, outperforming the dictionary-based system by more than 10%. This would imply that statistical word frequency patterns among dialects constitute a stronger signal to consider than mere lexical matching. Because of the high overlap (~70%) between dialect vocabularies, the machine learning system seems to have picked up on subtle distinctions in either word usage or word frequency that the rule-based method cannot leverage.

Interestingly, the relatively moderate performance (i.e., not beyond 80%) shows difficulties presented by intra-dialect similarities and informal written forms. These results mirror prior ones (e.g., in studies on the DART corpus), which showed difficulties in classifying Iraqi dialects even when considering them as one class. Table 2 summarizes the two models' performance and strength on the dataset, highlighting the pros and cons of each model when dealing with dialect classification tasks.

Table 2. Strengths and weaknesses of the models' performance on the dataset

Feature	Dictionary-Based Model	TF-IDF + SVM Model
Interpretability	Higher	Lower
Computational Efficiency	Higher	Lower
Accuracy	63.6%	74%
F1-score	63.4%	74%

5.2. Limitations of the study

The results can be considered very promising, but several limitations must be mentioned:

- Dialectal Overlap: About 70 percent of the lexical overlap generates classification problems, on an inherent basis, between dialects. A lot of words are common across different regions, especially when written.
- Ignoring Context: both models are bag-of-words-based, disregarding word order and meaning generated from the context, which may be crucial in revealing the dialectal variations.
- Informal Text Variability: Slang, abbreviations, and inconsistent spelling found in social media text samples tend to be a noise source. Advanced postprocessing could be investigated.
- Facebook Comments Only: The dataset may not represent the full spectrum of the target dialects. Other sources can be considered in future work by considering other platforms.

6. Conclusion

This work created a dataset from scratch, which included three regional dialects from Iraq, collecting samples from social media websites where the Iraqis communicate mostly. This study employed classification techniques using a conventional dictionary-based and machine learning-based model to verify the dataset's quality. The new dataset is a unique work that targets the audience in three regions in Iraq, which opens the doors for digital tools to analyse and process the Iraqi participation in social media platforms.

To prepare the dataset for model training and evaluation, the methodology included preparing and preprocessing a broad dataset of Iraqi text samples, including social media posts and user comments from selected regional categories. The dataset is divided into validation and training sets, allowing us to evaluate performance, adjust model parameters, and train the models without leading to overfitting.

The assessment metrics helped us determine each model's performance in recognizing Iraqi dialects from text data. The accuracy and F1-score metrics showed how the models' performance varied between the simple dictionary-based and SVM machine learning models. The metrics showed that the Dictionary-based model, due to its simplicity, showed the lowest performance, compared to the SVM model with the help of the TF-IDF vectorization method, which achieved superior results.

Acknowledgement

We thank Iran University of Science and Technology and the Department of Software, Computer Engineering College for their support and resources that were instrumental in this study.

References

[1] A Radford, Alec, et al. "Improving language understanding by generative pre-training." (2018).
[2] C Zhou, C Sun, et al. "A C-LSTM neural network for text classification," arXiv, Nov 2015, https://doi.org/10.48550/arXiv.1511.08630.
[3] A. Alnawas and N. Arici, "Sentiment Analysis of Iraqi Arabic Dialect on Facebook Based on Distributed Representations of Documents," ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol. 18, no. 3, Article 20, pp. 1–17, Sep. 2019, https://doi.org/10.1145/3278605.
[4] N. Tibi and M. A. Messaoud, "Arabic dialect classification using an adaptive deep learning model," Bull. Electr. Eng. Inform., vol. 14, no. 2, pp. 1108–1116, Apr. 2025, https://doi.org/10.11591/eei.v14i2.8165.
[5] Y. Matrane, F. Benabbou, and N. Sael, "A systematic literature review of Arabic dialect sentiment analysis," J. King Saud Univ. - Comput. Inf. Sci., vol. 35, no. 6, p. 101570, June 2023, https://doi.org/10.1016/j.jksuci.2023.101570.
[6] E Alsarsour, R Mohamed, and T. Elsayed, "DART: A Large Dataset of Dialectal Arabic Tweets," in Proc. 11th Int. Conf. Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 2018.
[7] A. Keleg and W. Magdy, "Arabic Dialect Identification under Scrutiny: Limitations of Single-label Classification," arXiv preprint, Oct. 2023, https://doi.org/10.48550/arXiv.2310.13661.
[8] I. Alansari, "Artificial Intelligence Model to Detect and Classify Arabic Dialects," J. Softw. Eng. Appl., vol. 16, pp. 287–300, Jul. 2023, https://doi.org/10.4236/jsea.2023.167015.
[9] A. Aliwy, H. Taher, and Z. AboAltaheen. (2020, Dec.). "Arabic Dialects Identification for All Arabic countries," Proc. Fifth Arabic Natural Language Processing Workshop [Online]. pp. 302–307. Available: https://aclanthology.org/2020.wanlp-1.32/.
[10] A. A. Hnaif, E. Kanan, and T. Kanan, "Sentiment Analysis for Arabic Social Media News Polarity," Intell. Autom. Soft Comput., vol. 28, no. 1, pp. 107–119, Feb. 2021, https://doi.org/10.32604/iasc.2021.015939.
[11] T. Kanan et al., "A Review of Natural Language Processing and Machine Learning Tools Used to Analyze Arabic Social Media," 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), Amman, Jordan, 2019, pp. 622-628, https://doi.org/10.1109/JEEIT.2019.8717369.

[12] A. Alnawas and N. Arici, "Sentiment Analysis of Iraqi Arabic Dialect on Facebook Based on Distributed Representations of Documents," ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol. 18, no. 3, Article 20, pp. 1–17, Sep. 2019, https://doi.org/10.1145/3278605.
[13] U. Braga-Neto, Fundamentals of Pattern Recognition and Machine Learning. Cham, Switzerland: Springer, 2020.
[14] P. Dangeti. Statistics for machine learning. UK: Packt Publishing Ltd, 2017.
[15] Jo, T. "Machine learning foundations: Supervised, Unsupervised, and Advanced Learning. Cham: Springer International Publishing." 2021.
[16] D. A. Pisner and D. M. Schnyer, "Support vector machine," in Machine Learning, A. Mechelli and S. Vieira, Eds. Academic Press, 2020, pp. 101–121.
[17] J. Lilleberg, Y. Zhu, and Y. Zhang, "Support vector machines and Word2vec for text classification with semantic features," in 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCICC)*, Beijing, China, 2015, pp. 136–140.
[18] D. Jurafsky and J. H. Martin, "Vector Semantics and Embeddings" in Speech and Language Processing, draft, Jan. 12, 2025.
[19] D. E. Cahyani and I. Patasik, "Performance comparison of tf-idf and word2vec models for emotion text classification," Bull. Electr. Eng. Inform., vol. 10, no. 5, pp. 2780–2788, Sep. 2021, https://doi.org/10.11591/eei.v10i5.3157.
[20] Scikit Developers. (2025, January 1). Scikit-learn: Machine Learning in Python [Online]. Available: https://scikit-learn.org.