King Saud University
College of Computer and Information Sciences
Department of Information Technology
Second Semester 1446H – Winter 2025
IT469 – Natural Language Processing

IT 469

Arabic Delicates classification Proposal

Prepared by:

Student Name	Student ID
Rana Alsayyari	443200565
Wassayef Alkherb	443200459
Bashair Alsadhan	443200668
Rama Alshebel	443200929
Maryam Altuwaijri	443200235

Supervised by:

Dr. Abeer Aldayel

Table of Contents

Table of Figures

Table of Tables

1. Introduction

Arabic dialect classification remains a challenging problem in NLP due to the language's diglossic nature [1] and the vast linguistic variations across regions, cities, and countries [2]. Traditional datasets for Arabic NLP often rely on text from online exchanges, such as emails and social media posts, which may not fully capture the nuances of spoken dialects [3]. However, the introduction of SADA, a dataset sourced from TV show dialogues and provided by the Saudi Broadcasting Authority in collaboration with SDAIA, presents new opportunities for dialect analysis.

A unique feature of SADA is its division into different domains, such as comedy and drama, which allows for a deeper understanding of how dialect is used in various contexts. This project aims to explore the performance of different classification models in out-of-domain dialect classification. Specifically, we seek to determine whether these models can generalize across different text domains by capturing dialectal patterns effectively or if their performance is restricted to the domains they are trained on. This analysis will provide valuable insights into the robustness of dialect classification models and their ability to handle real-world linguistic diversity.

Recent studies show how difficult Arabic dialect classification is. In [7], the authors fine-tuned AraBERT (a BERT model trained on large Arabic text collections), on 20,398 tweets labeled with 18 different country-level dialects. Even with strong models and a good amount of data, the evaluation on 4,758 tweets achieved only a 32.63% F1-score, showing how challenging the task remains, especially with short and informal texts. Similarly, in [8], the authors pointed out that most available datasets are limited, with short texts and only a few dialects represented. They suggested that using longer and more varied texts from different domains could help models perform better and generalize more effectively.

2. Experiment Setup

This chapter describes the experimental design used to evaluate Arabic dialect classification under domain shifts. We use the Saudi Audio Dataset for Arabic (SADA), a large corpus of transcribed speech across multiple dialects and media genres. We begin by preparing and balancing the dataset to ensure fair comparison across dialects and domains. Specifically, we focus on the three most represented dialects, (Najdi, Hijazi, and Khaliji) and the three most populated categories (Comedy, Drama, and Kids) carefully balancing the number of samples across all combinations. After preprocessing the text using standard normalization and cleaning techniques, we extract features suitable for machine learning models.

Our methodology focuses on text classification, to predict a speaker's dialect from a processed text input. We first establish a baseline using a simple Logistic Regression model trained on TF-IDF features. Then, we implement a more advanced solution using fine-tuned BERT transformers, which can capture deeper semantic representations of Arabic text. Both models are trained and evaluated under in-domain and out-of-domain settings to test their generalization capabilities.

The experimental results will later highlight the extent to which training on a specific domain affects model performance when deployed on unseen categories

2.1 Dataset

While English Speech Recognition has achieved remarkable success, thanks to the abundance of available data, the same progress has not been mirrored in Arabic Speech Recognition (ASR). One major challenge lies in the limited amount of training data, particularly for Dialectal Arabic, which varies greatly across regions. This makes the release of SADA (Saudi Audio Dataset for Arabic) [5] a significant milestone for the Arabic NLP community. SADA is the result of a collaboration between SDAIA (Saudi Data and Artificial Intelligence Authority) [4] and SBA (Saudi Broadcasting Authority) [6], where SBA provided audio from 80 different TV shows, and SDAIA handled the transcription, preparation, and public release of the dataset. Although the dataset was last updated two years ago on the Kaggle platform, no official update schedule has been specified.

The dataset, consists of 253,166 rows and 15 columns. It contains transcribed Arabic speech segments from various media genres, aiming to predict the speaker's dialect based on training data from different categories. Figure 1 present sample of SADA dataset, it includes metadata such as FileName, ShowName, FullFileLength, SegmentID, SegmentLength, SegmentStart, SegmentEnd, SpeakerAge, SpeakerGender, SpeakerDialect, Speaker, Environment, GroundTruthText, ProcessedText, and Category.

	Unnamed: 0	FileName	ShowName	FullFileLength	SegmentID	SegmentLength	SegmentStart	SegmentEnd	SpeakerAge	SpeakerGender	SpeakerDialect	Speaker	Environment	GroundTruthText	ProcessedText	Category
0	0	batch_1/6k_SBA_100_0.wav	الجلف بقلل الندى	605.3	6k_SBA_100_0-seg_249_720-251_840	2.12	249.72	251.84	Elderly -- كبير في السن	Male	Najdi	Speaker1-متحدث	Clean -- نظيف	ووضّح كلامك يا منيت	ووضيح كلامك يا منيت	درامي
1	1	batch_1/6k_SBA_100_0.wav	الجلف بقلل الندى	605.3	6k_SBA_100_0-seg_252_700-255_300	2.60	252.70	255.30	Elderly -- كبير في السن	Male	Najdi	Speaker1-متحدث	Clean -- نظيف	ترى رأسي ما عاد يتحمّل المزايك	ترى راسي ما عاد يتحمل المزايك	درامي
2	2	batch_1/6k_SBA_100_0.wav	الجلف بقلل الندى	605.3	6k_SBA_100_0-seg_256_010-258_180	2.17	256.01	258.18	Elderly -- كبير في السن	Male	Najdi	Speaker2-متحدث	Clean -- نظيف	سلامة رأسك يا ابو مسامح	سلامة راسك يا ابو مسامح	درامي

Figure 1: Sample of SADA dataset

The dialect distribution (as shown in Figure 2) indicates that Najdi (94,611 rows) is the most frequent dialect, followed by multi-speaker recordings (52,501 rows), Hijazi (36,170 rows), and Khaliji (30,320 rows). Other dialects such as Modern Standard Arabic, Egyptian, Levantine, Yemeni, and Maghrebi appear in much smaller proportions. The category distribution (in Figure 3) highlights that the dataset is dominated by Comedy (71,782 rows), Drama (52,843 rows), and Children's content (42,400 rows), while some categories, like Tourism (871 rows) and Entertainment (315 rows), are underrepresented.

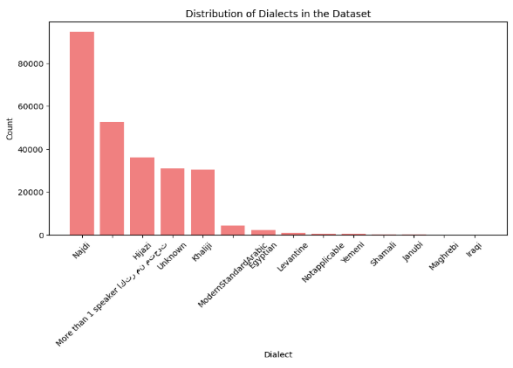

Figure 2: Distribution of dialects in the dataset.

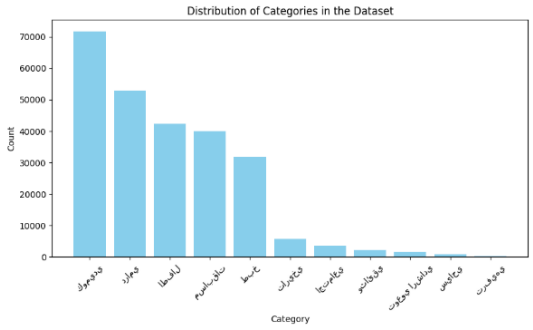

Figure 3: Distribution of categories in the dataset.

Figure 4 provides further insight into how different genres are distributed across dialects. It shows that some dialects appear more frequently in specific categories, which may influence the

training phase of the model. The selection of which categories to include in training will be determined in the next phase based on data distribution and relevance to the dialect classification task.

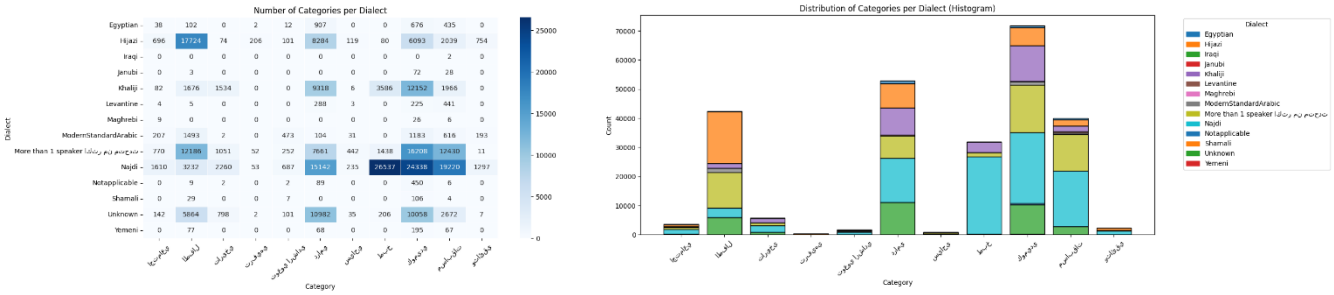

Figure 4: Distribution of categories per dialect.

Our question in this project is whether training a model on a specific category will affect its performance when tested on other categories (out-of-domain testing). In this experiment, we chose the top 3 dialects, Najdi, Hijazi, and Khaliji, and top 3 categories Comedy, Drama, and Kids.

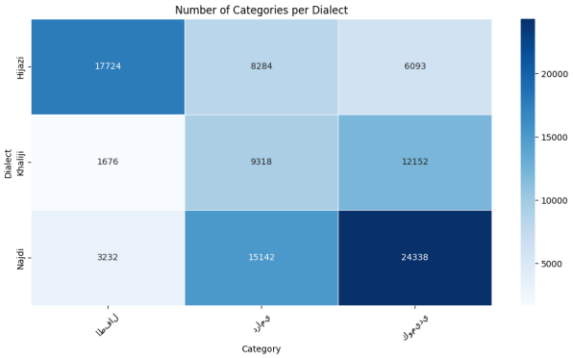

Figure 5: Dialect and categories distribution

Figure 5 shows the distribution of the chosen dialects and categories. To perform the out-of-domain testing, we will train 3 models, evaluate their performance, and test it on other categories. For fair comparison between the models, we will train and test them on equal rows. Figure 5 shows that the Khaliji dialect, kids category, has the least amount of data which is 1676 row, so we will select a random 1676 row for each dialect-category combination. Figure 6, show the final distribution of data after balancing it.

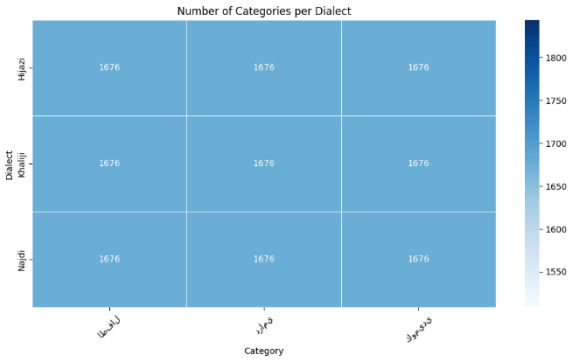

Figure 6: Dialect and categories balanced distribution

The dataset was originally published with a “ProcessedText” feature, where special characters, diacritics, emojis, and punctuations were removed. Additionally, Arabic letters were standardized into unified forms (e.g., آ، أ and إ were all normalized to ا), as stated in the Kaggle README [4]. Utterances and English words or letters were also removed during preprocessing. For our task, we only utilized the "ProcessedText" and "SpeakerDialect" features out of the 15 features in the dataset. Furthermore, the data was divided into three separate files based on their category. Additionally, we removed rows with missing values; this step resulted in the deletion of only four rows from the dataset. We used TF-IDF vectorization to transform the text data into numerical features for the Logistic Regression model. For BERT, we applied the BertTokenizer (bert-base-multilingual-cased), which tokenizes text into numerical token IDs that the model can understand.

2.2 Methodology

This section outlines the approach used to solve the task of Arabic dialect classification, with a particular focus on the model's ability to generalize across different text domains. The methodology centers on text classification, where the objective is to predict a dialect label (e.g., *Najdi*, *Hijazi*, *etc.*) from a transcribed Arabic utterance. We will train the model on a specific domain (such as *drama*) so that it can learn the dialectal patterns within that context. To evaluate the model's out-of-domain performance, we will then test it on data from a different domain (such as *comedy*). This setup allows us to examine how well the model can adapt to variations in linguistic style, register, and genre. We begin with a simple, interpretable baseline model using Logistic Regression, and gradually transition to more advanced transformer-based models, such as BERT. Through this process, we aim to understand the robustness, transferability, and limitations of dialect classification models when exposed to genre shifts.

Figure 7: Overview of the methodology used for Arabic dialect classification.

Describing the models used to achieve our goal, starting with a simple baseline and followed by an advanced transformer-based model for improved generalization, as follows:

Model Card 1: Logistic Regression (Baseline Model):

Model Details:

- **Developers**: Rana Alsayyari, Bashair Alsadhan
- **Model Date**: March 2025
- **Model Version**: First version (1.0)
- **Model Type**: Multinomial Logistic Regression (sklearn)
- **Training Algorithm**: Logistic Regression with L-BFGS solver
- **Hyperparameters**:
  - Penalty: l2 (default)
  - Solver: lbfgs
  - Max Iterations: 200
  - C (inverse regularization strength): 1.0

- **Features**: TF-IDF vectorization of the ProcessedText column (word ngrams 1–4, up to 5000 features)
- **Fairness Constraints**: N/A
- **Additional Approaches**:
  - Basic text preprocessing (normalization, diacritics removal, stemming, stop-word removal)
  - Balanced genre splits saved for cross-domain testing
  - TF-IDF vectorization
- **Paper/Resource**: Pedregosa, F. et al. (2011). *Journal of Machine Learning Research*, 12, 2825–2830 [1].
- **License**: N/A
- **Contact**: Rana Alsayyari, Bashair Alsadhan

Intended Use:

- **Primary intended uses**: Arabic dialect classification on pre-transcribed text (SADA dataset).
- **Primary intended users**: Researchers, students, and developers working on Arabic NLP or dialect detection
- **Out-of-Scope use cases**: Real-time or streaming audio, untranscribed speech, immediate production deployment

---

Metrics:

- **Performance measures**: Accuracy, precision, recall, F1-score, support (per class)
- **Decision thresholds**: Top-1 predicted class
- **Variation approaches**:
  - In-domain evaluation on the same genre
  - Cross-domain evaluation across comedy, drama, and kids
  - Analysis of performance drop across dialects and domains

Evaluation Data:

- **Dataset**: SADA Dataset (provided by SDAIA and SBA), we did three balanced CSVs (comedy, drama, kids), each split 80 % train and 20 % test
- **Motivation**: Benchmark cross-genre generalization for Arabic dialect classification
- **Preprocessing**:
    - Removal of duplicates and any records missing text or label.
    - Text normalization and diacritic removal.
    - TF-IDF transformation of ProcessedText column

Training Data:

- **Dataset**: 80 % train / 20 % test of the cleaned SADA data (combined or per-genre) **Distribution**: Dialect label distribution aligned with the original SADA distribution
- **Sampling**: Random shuffle with fixed seed for reproducibility

---

Quantitative Analyses:

- **In-domain results:** Test accuracy and classification report on each genre's test set.
- **Cross-domain results:** Accuracy and classification report when testing on held-out genres (e.g. train on comedy → test on drama/kids).
- **Intersectional analyses:** N/A

Note: See Section 3 – Evaluation and Results for all numeric values.

Ethical Considerations

- Trained solely on publicly sourced, media-transcribed text (no private or spontaneous speech).

Caveats and Recommendations

- Baseline model not tuned for production
- Limited cross-domain generalization expected
- Recommended as a benchmarking and academic baseline

Model Card 2: BERT Transformer (Advanced transformer-based model):

Model Details:

- **Developers**: Rama Alshebel, Wassayef Alkherb and Maryam Altuwaijri
- **Model Date**: April 2025
- **Model Version**: First version (1.0)
- **Model Type**: Multilingual BERT (Bidirectional Encoder Representations from Transformers)
- **Pretrained Checkpoint:** Base multilingual cased BERT
- **Training Algorithm**: Fine-tuning with weight-decay-regularized AdamW and linear learning-rate decay
- **Hyperparameters**:
    - Number of classes: 3
    - Maximum sequence length: 128 tokens
    - Batch size: 16
    - Epochs: 10
    - Learning rate: 2×10^{-5}
    - Weight decay: default
    - LR schedule: linear decay, zero warmup
- **Features**:
    - Text tokenized into word-piece subword units
    - Padded or truncated to 128 tokens
    - Input field: cleaned ProcessedText
- **Fairness Constraints**: N/A
- **Additional Approaches**:
    - Label encoding of dialect classes
    - Genre-balanced sampling (comedy, drama, kids)
    - Cross-genre evaluation (train on one domain, test on others)
- **Paper/Resource**: Devlin, J. *et al.* (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.*
- **License**: N/A
- **Contact**: Rama Alshebel, Wassayef Alkherb and Maryam Altuwaijri

---

Intended Use:

- **Primary intended uses**: Arabic dialect classification on pre-transcribed text (SADA dataset).
- **Primary intended users**: Researchers, students, and developers working on Arabic NLP or dialect detection
- **Out-of-Scope use cases**: Real-time or streaming audio, untranscribed speech, immediate production deployment

---

Metrics:

- **Performance measures**: Accuracy, precision, recall, F1-score, support (per class)
- **Decision thresholds**: Top-1 predicted class
- **Variation approaches**:
    - In-domain evaluation on the same genre
    - Cross-domain evaluation across comedy, drama, and kids
    - Analysis of performance drop across dialects and domains

---

Evaluation Data:

- **Dataset**: SADA Dataset (provided by SDAIA and SBA), we did three balanced CSVs (comedy, drama, kids), each split 80 % train and 20 % test
- **Motivation**: Benchmark cross-genre generalization for Arabic dialect classification
- **Preprocessing**:
    - Removal of duplicates and any records missing text or label.
    - Label-encode the dialect classes
    - Tokenize into word-piece subword units, pad/truncate to 128 tokens

---

Training Data:

- **Dataset**: 80 % train / 20 % test of each genre-specific balanced SADA subset
- **Distribution**: Equal representation of the three dialects within each genre
- **Sampling**: Random shuffle with fixed seed for reproducibility

---

Quantitative Analyses:

- **In-domain results:** Test accuracy and classification report on each genre's test set.
- **Cross-domain results:** Accuracy and classification report when testing on held-out genres (e.g. train on comedy → test on drama/kids).
- **Intersectional analyses:** N/A

Note: See Section 3 – Evaluation and Results for all numeric values.

Ethical Considerations

- Trained solely on publicly sourced, media-transcribed text (no private or spontaneous speech).

Caveats and Recommendations

- **Computer requirements:** High (GPU recommended for fine-tuning and inference)
- **Generalization:** May degrade on noisier or entirely new domains
- **Deployment readiness:** Requires further tuning, validation, and domain adaptation before production use.

Beyond standard supervised training, we frame our cross-genre evaluation as a form of transductive transfer learning, specifically domain adaptation. In practice, we train each model (TF-IDF + logistic regression or fine-tuned BERT) on data from one or more source genres (e.g. comedy) and then directly apply it to held-out target genres (drama, kids). This setup lets us measure how well each feature representation and learning algorithm transfers dialectal classification knowledge across distinct domains.

3. Evaluation and Results

Tables 1 and 2 present the performance of the Logistic Regression and BERT models across different training and testing combinations using key evaluation metrics: Accuracy (A), Precision (P), Recall (R), and F1-score (F1).

In real-world scenarios such as dialect identification, **F1-score** is particularly important, as it balances Precision (how many predicted dialects were correct) and Recall (how many dialects were successfully detected). This is critical when both false positives and false negatives carry impact based on their application. For instance, misidentifying a dialect could affect tasks like text-to-speech, targeted content delivery, or region-specific NLP services.

Table 1: Logistic Regression model result.

	Test - Comedy	Test - Drama	Test - kids
Train-Comedy	**A: 0.405** P: 0.41 R: 0.41 F1: 0.40	**A: 0.349** P: 0.35 R: 0.35 F1: 0.35	A: 0.351 P: 0.35 R: 0.35 F1: 0.35
Train-Drama	**A: 0.347** P: 0.35 R: 0.35 F1: 0.34	**A: 0.422** P: 0.42 R: 0.42 F1: 0.42	**A: 0.360** P: 0.36 R: 0.36 F1: 0.35
Train-Kids	A: 0.338 P: 0.33 R: 0.34 F1: 0.33	A: 0.348 P: 0.35 R: 0.35 F1: 0.34	**A: 0.494** P: 0.49 R: 0.49 F1: 0.49

Table 2: Bert model results

	Test - Comedy	Test - Drama	Test - kids
Train-Comedy	**A: 0.395** P: 0.40 R: 0.40 F1: 0.40	A: 0.341 P: 0.35 R: 0.35 F1: 0.35	**A: 0.361** P: 0.35 R: 0.35 F1: 0.35
Train-Drama	**A: 0.397** P: 0.39 R: 0.39 F1: 0.39	**A: 0.407** P: 0.39 R: 0.40 F1: 0.39	A: 0.334 P: 0.37 R: 0.37 F1: 0.36
Train-Kids	A: 0.348 P: 0.33 R: 0.34 F1: 0.33	**A: 0.361** P: 0.35 R: 0.35 F1: 0.34	**A: 0.472** P: 0.49 R: 0.49 F1: 0.49

4. Discussion

This section will discuss the results of cross-domain training obtained from logistic regression and BERT by comparing their accuracy.

In our journey to investigate a model's ability to recognize Arabic dialects across different domains, we developed two models: Logistic Regression and BERT model.

Both models were trained to identify whether a speaker used Hijazi, Khaliji, or Najdi dialects, with crossing domains. The training was done on a single domain (for example, Comedy), and the testing was done on either the same or completely different domains (Kids or Drama).

This design gave us a chance to examine a fundamental question:

If a model is trained within the boundaries of one domain, can it successfully adapt to a new and different linguistic domain?

The patterns for logistic regression were obvious and convincing, as shown in Table 1. When the model was trained and tested within the same domain, it had reasonable outcomes, with an accuracy of 40.5% on Comedy and 42.2% on Drama. However, after testing on a new and different domain, the model struggled to maintain its performance. It is slightly dropped when crossing into another domain, indicating that Logistic Regression was sensitive to shifts in vocabulary and tone.

Despite some inconsistencies, one domain had a more precise and stable trend: kids. Whether the kid’s data was trained or tested, Logistic Regression achieved higher performance. When trained and tested on Kids' data, it had a high score of almost 49.4% accuracy, which was the highest across all domains.

This highlights a crucial linguistic insight: kids' content is generally more simplistic, with higher repetition, lower ambiguity, and a clearer dialect and language pattern, thus making it easier for models to learn within and generalize out of it.

After developing logistic regression, we moved to BERT, which provides a deeper understanding of language and context. While BERT was initially expected to generalize across domains seamlessly, the results showed the opposite. As shown in Table 2, BERT performed comparably, with an accuracy of 39.9% on Comedy and 39.7% on Drama. Despite its

unpredictable architecture, BERT struggled when crossing domains, as we previously noticed with Logistic Regression. However, as we noticed with Logistic Regression, BERT succeeded in the Kids' domain, achieving 48.2% with a trained and tested Kids' data.

What both results fundamentally revealed from these two entirely different models was something meaningful:

The challenge lies not in the model's architecture but in the domain's nature.

The Kids domain has relative constraints and regularity in linguistics, predictable patterns, simpler words, and easier accessibility to interpretation.
In comparison, Comedy and Drama contain offense, emotion, and non-standard linguistic forms, all of which add complexity and make it harder for models to decode.

While some of the variation in performance can be attributed to model architecture and domain complexity, the data can also be a significant factor! In particular, inconsistency in data labeling can disrupt the model's learning and, therefore, its generalizing ability. Inconsistent or wrong labels create noise in the data, which causes it to be more difficult for the model to learn meaningful patterns in the data and generalize well. Upon inspecting the dataset, we found some mislabeled examples, which might have impacted model performance on the test data. As seen in Table 3, several examples show instances where the text did not match the category label. For example, a few examples were classified as Najdi, where they are Hijazi, and vice versa.
Such misclassification adds label noise and directly affects the model's ability to learn appropriate pattern distinctions, resulting in lower accuracy.

Table 3: Examples of Misclassified Dataset Entries

Category	Processed Text	Speaker Dialect
طالع في صحتك يا بني ادم يا بني ادم كل بالهنا والشفا يا تاكل يا تقول شبعت وتقوم	Drama	Najdi

ما لهم دول كده نايمين طب وبعدين يا ام الخير الخيمة ما فيها اكل بالمره انت ايش حتسوي دول نايمين ولا يصحوا لا بكرة ولا بعد بكرة يمكن ام الخير متفضل جوعانة كده على طول وكمان بابا فرحان وسندباد ما هم موجودين	Kid	Najdi
المسلسل الحاي ما في الا مقاطع بلا لا تستحون	Comedy	Hijazi
يا الله حي الشباب وراكم مستحين تقضلوا انا الا حياكم	Comedy	Hijazi

We also assessed our models using the same sentences to evaluate better and characterize the models' behavior differences across domains. Since we observed that the Kid model achieved a higher performance, we evaluated this model in different domains. As Table 4 shows, the first sentence from the Comedy domain, "طيب يا مناحي هذا اني نازلة", although originally Najdi, was predicted as Hijazi by both models. Interestingly, in the case of the Drama domain, the sentence, "وكان يضحك عليا" was confused in different ways with the Logistic Regression predicting Najdi, while BERT predicted Khaliji. In the Kids domain, the sentence, "انا اوريكي يا بقرة" produced two uniquely different predictions: BERT predicted Hijazi while Logistic Regression predicted Khaliji.

The evaluation example provides additional evidence in favor of our broader takeaways. In some cases, the two models often had very different predictions based on the domain context and model architecture, even with the same sentences. This reinforces the notion that cross-domain generalization is complex, especially as linguistic indications in dialects may be nuanced, but their shifts with genre or context can be drastic. Although BERT, as a more robust linguistic model than the Logistic Regression models, was still consistent within their training domain. Overall, the examples demonstrate that the training domain provides the models a more explicit "linguistic map", thus allowing for almost instant recognition of expected patterns.

Table 4: Model Predictions Across Domains on same sentences

	Sentence	Actual value	Predicted value-LR	Predicted value-Bert
Test on Comedy	طيب يا مناحي هذا اني نازلة	Najdi	Hijazi	Hijazi
Test on Drama	وكان بينضحك عليا	Hijazi	Najdi	Khaliji
Test on Kids	انا اوريكي يا بقرة	Hijazi	Khaliji	Hijazi

McNemar's Test:

In this section, we will present and analyze the cross-domain evaluation results for the two models by applying McNemar's Test.

Table 1 and Table 2 present the cross-domain evaluation of two different models, where each model was trained on a single category (Comedy, Drama, or Kids) and then tested across all three categories. As expected, each model performs best on the category it was trained on, for example, the model trained on Drama achieves the highest accuracy when tested on Drama. When tested on the other two categories, performance drops slightly but remains relatively close.

To investigate whether training on a specific category actually leads to a statistically significant improvement in performance, we applied McNemar's Test, a statistical method used to compare two models (within logistic regression or Bert) on the same test set. McNemar's Test doesn't just look at overall accuracy; it digs deeper by comparing the individual predictions of the two models. For each data point, it checks whether both models got it right, both wrong, or whether one succeeded where the other failed. This allows us to identify real differences in model behavior, even when their overall accuracies are similar.

For instance, two models might achieve nearly identical accuracy scores, yet one might be better at predicting the Hijazi dialect, while the other excels at Najdi. Relying on accuracy alone

might mask these differences. McNemar's Test helps uncover these subtleties and gives us a more reliable sense of whether one model truly outperforms the other in practice.

We performed McNemar's Test three times, once for each test category (Comedy, Drama, and Kids). For each test, we compared two models:

1. The model trained on the same category as the test set (e.g., Comedy model tested on Comedy).
2. The highest-performing model from the other categories on that same test set.

For example, based on Table 1, when evaluating on the Comedy test set, we compared the Comedy-trained model with the Drama-trained model, since it performed better than the Kids model on that test set. We followed the same logic for Drama and Kids, always comparing the in-domain model (trained on the same category) with the strongest cross-domain alternative. To make it easy to track, we have bolded in the table the models that were selected for each test comparison.

- **McNemar's Test LR:**

In this section, we will present and analyze the cross-domain evaluation results for the LR model by applying McNemar's Test.
Let's take a closer look at the results for each test set individually in LR:

Comedy Test

Figure 8, shows the results of the McNemar's Test performed on the Comedy test set. The p-value was 0.007, which is less than 0.05 -the common threshold for statistical significance-, indicating the model trained on Comedy behaves significantly differently than the one trained on Drama when classifying Comedy data.

This result suggests that the Drama-trained model does not generalize well to the Comedy domain, since the model trained directly on Comedy performs significantly differently and better when tested on Comedy data. In other words, training on a different category (like Drama) leads to meaningful changes in performance when predicting Comedy, indicating a domain mismatch.

```
Contingency Table: [[0, 256], [198, 0]]
P-value: 0.007402294925029015
The difference between the two baselines is statistically significant.
```

Figure 8: McNamer test results on Comedy Category

Drama Test

Figure 9 presents the results of the McNemar's Test conducted on the Drama test set. The p-value was 0.00056, which is well below the 0.05 threshold for statistical significance.

This result shows that the Comedy-trained model behaves quite differently from the Drama-trained model when both are tested on Drama data. The Drama-trained model performs better, and this difference isn't just by chance, it's statistically significant. In other words, a model that was trained directly on Drama clearly understands Drama better. This highlights an important point: models trained on one category don't always work well on another, and switching domains can lead to noticeable drops in performance.

```
Contingency Table: [[0, 262], [188, 0]]
P-value: 0.0005642764050526149
The difference between the two baselines is statistically significant.
```

Figure 9: McNamer test results on Drama Category

Kids Test

Figure 10 shows the results for the Kids test set. The p-value from the McNemar's Test is incredibly small (1.5e-10), which is far below the 0.05 threshold. This means the difference between the two models is statistically significant.

In this case, the Kids-trained model and the Drama-trained model gave noticeably different predictions when tested on Kids data, and the Kids-trained model did better. So, even though both models were evaluated on the same data, their predictions weren't the same, and this difference matters.

This result tells us that a model trained specifically on Kids data captures its patterns much more effectively, while models trained on other categories (like Drama) may not generalize well to this domain.

```
Contingency Table: [[0, 262], [188, 0]]
P-value: 0.0005642764050526149
The difference between the two baselines is statistically significant.
```

Figure 10: McNamer test results on Kids Category

The main takeaway from these tests is: Training on the same category as the test set leads to better-aligned predictions and improved performance.

All three McNemar's Tests, for the Comedy, Drama, and Kids test sets, showed statistically significant differences between the in-domain model and the best-performing out-of-domain model. This means that, across all categories, the model trained specifically on the target domain consistently made different and more accurate predictions compared to models trained on other domains.

This confirms that domain-specific training has a real impact, and that models may struggle to generalize well across different categories, even when accuracy scores seem close.

- **McNemar's Test Bert:**

In this section, we will present and analyze the cross-domain evaluation results for the BERT model by applying McNemar's Test.

Comedy Test

Figure 11 shows the results of McNemar's Test performed on the Comedy test set. Since this p-value is greater than the common significance threshold of 0.05, we conclude there is no statistically significant difference between the two models' behaviors. Although the two models disagreed on several individual predictions, the difference between their overall performances is insignificant enough to be considered meaningful. This suggests that both models generalize similarly to the Comedy domain, even though they were trained in different genres (Comedy vs. Drama).

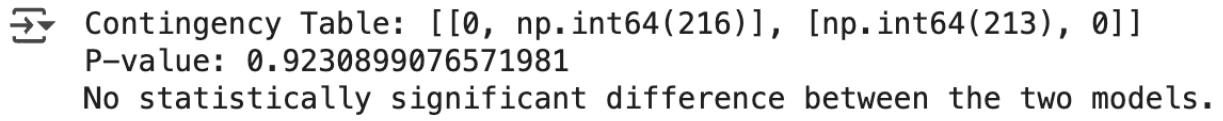

Figure 11: McNemar's test results on Comedy category (BERT)

Drama Test

Figure 12 shows the results of McNemar's Test performed on the Comedy test set. Since the p-value is less than 0.05, we conclude that there is a statistically significant difference between the two models' behaviors when evaluated in the Drama domain. This means that the differences we observe in their predictions are unlikely to be due to random chance. Although both models attempt to classify the dialects within drama texts, their underlying learned patterns and generalization strategies differ enough to produce distinct prediction outcomes. This suggests that the models approach dialectal features in the Drama genre differently, likely influenced by the linguistic characteristics of their respective training domains. In real-world terms, this tells us that the training genre has a meaningful impact on how well the model adapts to the drama domain.

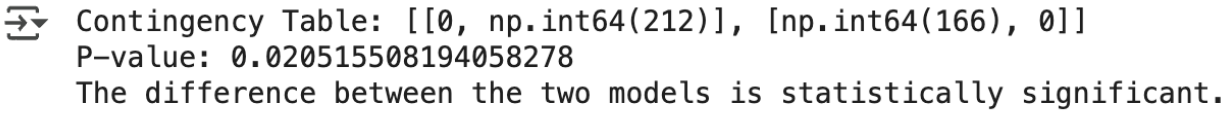

Figure 12: McNemar's test results on Drama category (BERT)

Kids Test

Figure 13 shows the results of McNemar's Test performed on the Comedy test set. Since the p-value is far below 0.05, we again conclude that there is a statistically significant difference between the two models' behaviors when evaluated in the Kid's domain. In this case, the significant disparity between the two models further emphasizes that the genre in which a model is trained can heavily influence its ability to generalize to new types of data, particularly in domains like Kids, where language style and structure are quite distinct.

```
Contingency Table: [[0, np.int64(313)], [np.int64(188), 0]]
P-value: 2.5783454862818024e-08
The difference between the two models is statistically significant.
```

Figure 13: McNemar's test results on Kid category (BERT)

McNemar's Test repeatedly indicated that domain-specific training caused the models to make statistically different predictions compared to cross-domain models in most cases. Although not every comparison achieved statistical significance, the general trends indicate that domain-specific training enhances alignment and performance. Training in the same category consistently provides more aligned predictions. Even with deep models like BERT, cross-domain generalization remains a challenge.

Domain-specific training increases the accuracy and consistency of prediction, not only with overall accuracy metrics. While the BERT models are likely much more powerful than the logistic regression models, they still struggle when presented with new, unseen linguistic contexts. This indicates that data with more domain diversity is critical in training models.

5. Conclusion

In this project, our primary objective was to explore how well Arabic dialect classification models could generalize across different domains when trained on a specific category. To achieve this, we conducted cross-domain evaluations by training models on one domain (Comedy, Drama, or Kids) and testing them on the same and different domains. We experimented with two models: Logistic Regression, as a baseline, and BERT, as a more advanced language representation model.

Our results showed that both models performed best when trained and tested within the same domain, highlighting the impact of domain-specific characteristics on model performance.

- **Logistic Regression** achieved its highest accuracy of 49.4% when trained and tested on the Kids domain, and an accuracy of 40.5% and 42.2% when trained and tested on Comedy and Drama respectively.
- **BERT** similarly performed best on the Kids domain with an accuracy of 47.2%, and achieved 39.5% and 40.7% on Comedy and Drama respectively.

Cross-domain testing revealed noticeable performance drops, confirming that linguistic differences across domains challenge both simpler and complex models alike.

Further statistical analysis using McNemar's Test supported these observations:

- **For Logistic Regression**, statistically significant differences were found in all three test categories: Comedy ($p = 0.007$), Drama ($p = 0.00056$), and Kids ($p = 1.5e-10$).
- **For BERT**, statistically significant differences appeared in Drama and Kids (Drama test $p< 0.05$, Kids test $p < 0.05$), but not in Comedy.

These results reveal that both models despite their architectural differences struggled with cross-domain generalization. Performance consistently dropped when a model trained on one domain was evaluated on another. However, both models achieved notably better results when working within the Kids domain, which suggests that linguistic simplicity, predictability, and less ambiguous structure significantly aid model learning and transferability.

Further analysis using McNemar’s Test strengthened these observations, confirming that models trained on the same domain as the test data performed statistically differently than those trained on different domains. Even BERT, despite its deep contextual understanding, faced challenges when dealing with domain shifts, emphasizing that the complexity and inconsistency across domains pose significant hurdles to generalization.

Several factors, such as domain-specific vocabulary, linguistic complexity, and even mislabeling in the dataset, were identified as contributing to performance variability. Mislabeling, in particular, introduced noise that disrupted the models' ability to learn clean patterns, further highlighting the importance of high-quality, accurately labeled data.

For future work, improvements could include:

- Using a more consistent and accurately labeled dataset to reduce label noise.
- Investigating larger and more diverse pre-trained models fine-tuned specifically for Arabic dialect identification tasks.

Overall, our findings underscore that domain-specific characteristics strongly influence model performance, and tackling domain shifts remains a key challenge in building reliable, real-world NLP systems.

6. References

List references using a proper style such as IEEE

[1]L. A. Al Suwaiyan, “Diglossia in the Arabic Language,” International Journal of Language & Linguistics, vol. 5, no. 3, 2018, doi: https://doi.org/10.30845/ijll.v5n3p22.

[2] A. A. Alsuwaylimi, “Arabic Dialect Identification in Social Media: A Hybrid Model with Transformer Models and BiLSTM,” Heliyon, vol. 10, no. 17, pp. e36280–e36280, Aug. 2024, doi: https://doi.org/10.1016/j.heliyon.2024.e36280.

[3] https://direct.mit.edu/coli/article/40/1/171/1458/Arabic-Dialect-Identification O. F. Zaidan and C. Callison-Burch, “Arabic Dialect Identification,” Computational Linguistics, vol. 40, no. 1, pp. 171–202, Mar. 2014, doi: https://doi.org/10.1162/coli_a_00169.

[4] SADA صدى, https://www.kaggle.com/datasets/sdaiancai/sada2022/data?select=test.csv Accessed: 2025-02-27

[5] S. Alharbi, A. Alowisheq, Z. Tüske, K. Darwish, A. Alrajeh, A. Alrowithi, A. Tamran, A. Ibrahim, R. Aloraini, R. Alnajim, R. Alkahtani, R. Almuasaad, S. Alrasheed, S. Alsubaie, and Y. Alonaizan, “SADA: Saudi Audio Dataset for Arabic,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), 2024, pp. 10286–10290, doi: 10.1109/ICASSP48485.2024.10446243.

[6] “هيئة الاذاعة والتلفزيون,” *Sba.sa*, 2021. https://sba.sa/ar

[7] J. Attieh and F. Hassan, "Arabic Dialect Identification and Sentiment Classification using Transformer-based Models," in *Proc. 7th Arabic Natural Language Processing Workshop (WANLP)*, Abu Dhabi, United Arab Emirates (Hybrid), Dec. 2022, pp. 485–490, doi: 10.18653/v1/2022.wanlp-1.54.

[8] M. Abdul-Mageed, A. Keleg, A. Elmadany, C. Zhang, I. Hamed, W. Magdy, H. Bouamor, and N. Habash, "NADI 2024: The Fifth Nuanced Arabic Dialect Identification Shared Task," in *Proc. 2nd Arabic Natural Language Processing Conference*, Bangkok, Thailand, Aug. 2024, pp. 709–728, doi: 10.18653/v1/2024.arabicnlp-1.79.