SADSLyC: A Corpus for Saudi Arabian Multi-dialect Identification through Song Lyrics

Salwa Alahmari[1,3], **Eric Atwell**[1], **Mohammad Alsalka**[1] **and Hadeel Saadany**[2]
[1]University of Leeds, UK,
[2]Birmingham City University and
[3]University of Hafr Al Batin, Saudi Arabia
{scssala, e.s.atwell, m.a.alsalka}@leeds.ac.uk
hadeel.saadany@bcu.ac.uk

Abstract

This paper presents the Saudi Arabian Dialects Song Lyrics Corpus (SADSLyC), the first dataset featuring song lyrics from the five major Saudi dialects: Najdi (Central Region), Hijazi (Western Region), Shamali (Northern Region), Janoubi (Southern Region), and Shargawi (Eastern Region). The dataset consists of 31,358 sentences, with each sentence representing a self-contained verse in a song, totaling 151,841 words. Additionally, we present a baseline experiment using the SaudiBERT model to classify the fine-grained dialects in the SADSLyC Corpus. The model achieved an overall accuracy of 73% on the test dataset.

1 Introduction

Through the analysis of Arabic song lyrics, one can explore the rich linguistic nuances of the Arabic language, recognise regional variations, and appreciate the artistic and literary elements present in the music. Within the structure of a song, verses often serve as the storytelling components, unravelling the plot or message, while choruses provide a recurring, emotive anchor that reinforces the song's central theme[1].

The lyrics of Arabic songs available online are categorised based on the singer's country of origin, regardless of the actual dialect of the lyrics. Taking Nancy Ajram, a famous Arabic singer, as an example[2], despite frequently singing in the Egyptian Arabic dialect, she is consistently recognised as a Lebanese singer. Thus, whether her songs are in Lebanese Arabic or not, they are invariably placed within the list of Lebanese songs on any musical platform (El-Haj, 2020). In this study, we do not rely on this classification to identify the Saudi Arabian dialect of the song lyrics. Instead, we used the songwriter's region of origin to construct the SADSLyC Corpus. We believe this approach is more accurate since it focuses on the lyrics written by the songwriter rather than the singer, who simply performs what is written.

The SADSLyC corpus consists of 1,892 Saudi Arabian songs, encompassing 31,358 sentences, and representing the five primary Saudi Arabian dialects: Najdi, Hijazi, Shamali, Janoubi,and Shargawi. In this paper, we will use the terms sentence and verse interchangeably.

The structure of this paper is as follows: Section 2 outlines related works. Section 3 describes the research methodology of this study. Section 4 provides a description of the SADSLyC corpus along with its statistical details. Section 5 presents the baseline experiment and results. Section 6 discusses the implications of the findings, addresses the limitations of the study. Finally, Section 7 provides a conclusion and suggestions for future work.

2 Related Work

The Habibi Corpus, developed by El-Haj (2020), is currently the only Arabic song lyrics dataset available in the literature. This corpus comprises 30,000 Arabic songs from 18 countries, covering six Arabic dialects: Egyptian, Levantine, Gulf, Maghrebi, Iraqi, and Sudanese. For dialect identification, the corpus was automatically labeled based on the nationality of the singer, providing a foundational resource for Arabic dialectal analysis in song lyrics.

Specific to the Saudi Arabian dialect, much of the prior research has focused on sentiment and emotion analysis in Saudi social media, particularly Twitter. Studies by AlMazrua et al. (2022), Almuqren and Cristea (2021), and others, such as Al-Twairesh et al. (2018), AL-Rubaiee et al. (2017), and Assiri et al. (2016), have provided valuable insights into this area. Additionally, Bayazed et al. (2020) classified Saudi tweets according to subdialects and sentiment, advancing the study of lin-

[1] https://en.wikipedia.org/wiki/Songstructure
[2] https://en.wikipedia.org/wiki/NancyAjram

Proceedings of The 4th Workshop on Arabic Corpus Linguistics (WACL-4), pages 38–43
January 20th, 2025. ©2024 Association for Computational Linguistics

guistic and emotional nuances within Saudi Arabic.

However, none of the previously mentioned studies focus on fine-grained dialects of Saudi Arabia based on geographical location. In our previous work (Alahmari et al., 2024), we employed Twitter for the Arabic dialect identification task, using ChatGPT for the identification process. We collected a small dataset from Twitter using dialectal word lists, representing the five main dialects of Saudi Arabia: Najdi, Hijazi, Shamali, Janoubi, and Shargawi.

The SADSLyC corpus stands distinct from existing literature due to its focus on a new genre, specifically Saudi song lyrics. As noted by Almuqren and Cristea (2021), the majority of Saudi corpora have primarily relied on Twitter as the sole data source.

3 Methodology

This section provides details on the construction of the SADSLyC corpus, including the selection of songwriters, data collection, data preprocessing, and data labeling.

3.1 Data Selection Criteria

Initially, we dedicated a considerable amount of time to seeking out Saudi songwriters hailing from diverse regions, representing the five primary dialects relevant to our study. Our approach to gathering information about the hometown or birthplace of each songwriter involved leveraging two main web-based resources: Wikipedia[3] and Google[4]. Typically, Wikipedia provides details about the hometown or birthplace of the songwriter. However, there were instances where the Wikipedia page for the songwriter did not exist. Furthermore, in other cases, essential information regarding the hometown or birthplace was absent. Consequently, we extended our search to include web pages such as forums, blogs, and Twitter accounts in pursuit of information about the songwriter. Additionally, we delved into YouTube, scouring TV interviews that shed light on the songwriter's hometown or birthplace. When the necessary information remained elusive from the aforementioned sources, we resorted to investigating the origin of the songwriter's family. Notably, many family names in Saudi Arabia correspond to renowned tribal names, particularly in the southern (Janoub) and northern (Shamal) regions, exemplified by well-known tribes like Alqahtani and Alshammari. Finally, when we were unable to find specific information regarding the songwriter's hometown or birthplace, we excluded the song from the list.

3.2 Data Collection

For the data (song lyrics) collection, we utilized the Web as Corpus method (Kilgarriff and Grefenstette, 2001). There are a large number of websites that provide textual representations of song lyrics. However, not all of them provide information about the songwriter or allow web scraping techniques.

We primarily extracted song lyrics (textual data) from three web sources: Wneen[5], Kalimat Aghani[6], and Fnanen[7]. After inspecting the HTML pages of each website, we developed Python code using the BeautifulSoup4 library[8] to scrape the website based on its HTML elements.

3.3 Data Preprocessing

To ensure the SADSLyC corpus is free from unwanted elements such as advertisements, spam, hashtags, or symbols, we implemented preprocessing and data-cleaning methods. This meticulous approach results in a refined corpus devoid of any noise. To achieve this, we utilised the arabicprocess[9] library in Python for cleaning and preprocessing Arabic text.

3.4 Data Labeling

As previously mentioned, dialect labels are assigned to the lyrics based on the songwriter's origin. For instance, renowned Saudi poet and songwriterخالد الفيصل Khalid Alfaisal[10] originates from Najd (central Saudi Arabia) and resides in Riyadh[11]. Consequently, all songs authored by him are labeled as "Najdi". Similarly, songs penned by ثريا قابل Thuraya Qabel[12], a Saudi songwriter from Hijaz (western Saudi Arabia) who resides in Jeddah[13], are labeled as "Hijazi".

As the final step, these lyrics were assigned to two native speakers from each of the five dialect regions, totaling 10 native speakers. They validated

[3] www.wikipedia.org/

[4] www.google.com/

[5] https://www.wneen.com

[6] https://www.kalimataghani.com

[7] https://fnanen.com

[8] https://pypi.org/project/beautifulsoup4/

[9] https://pypi.org/project/arabicprocess/

[10] https://en.wikipedia.org/wiki/KhalidAl-Faisal

[11] https://en.wikipedia.org/wiki/Riyadh

[12] https://en.wikipedia.org/wiki/ThurayaQabil

[13] https://en.wikipedia.org/wiki/Jeddah

the labels and ensured that the lyrics accurately represented their respective dialects.

4 Corpus Description

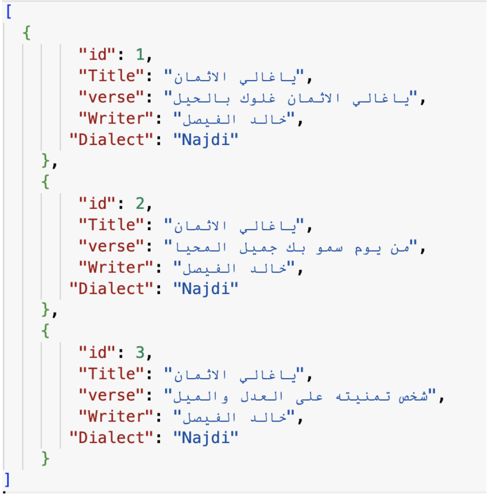
The original song lyrics are parsed into sentences based on the verses. The finalized corpus is saved in JSON format. Each song verse is assigned a unique "id" number, with the verse content stored under the "verse" field. Verses belonging to the same song are associated with the same title, writer, and dialect. Figures 1,2,3,4, and 5 show samples from the SADSLyC corpus JSON files for Najdi, Hijazi, Shamali, Janoubi, and Shargawi, respectively.

The corpus is available[14] for academic and research purposes to enrich the development of Arabic linguistic resources.

Sub-Dialect	Sentence Count	%	#Songs
Najdi	19481	62.12%	1118
Hijazi	7359	23.47%	392
Janoubi	1960	6.25%	129
Shamali	1017	3.24%	110
Shargawi	1541	4.91%	143
Total	31358	100%	1892

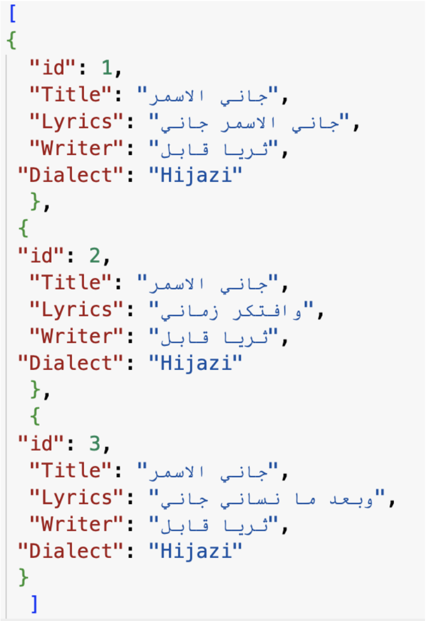
Table 1: The SADSLyC Corpus Sentence Count by Dialect

The corpus statistics in Table 1 clearly show that Najdi songs make up a significant portion of the corpus, accounting for 64.52%. The high percentage of Najdi songs can be attributed to several factors. **Firstly**, the dominance of the Najdi dialect in Saudi songs plays a significant role, as many well-known Saudi songwriters originate from Najd, further contributing to this prevalence. **Secondly**, our search for Saudi songwriters from the five regions of Saudi Arabia revealed that Shargawi songwriters tend to write in MSA rather than in the Saudi Arabian dialect, which has resulted in a limited collection of song lyrics in the Shargawi dialect. Additionally, poets from the Janoubi and Shamali regions prefer to compose Shilaat, a unique style of song that is typically performed without music. However, written sources for Shilaat are scarce, as most of them are available online in video or audio format. Consequently, we have a smaller portion of Janoubi and Shamali textual song lyrics in our corpus.



```
[
  {
      "id": 1,
      "Title": "ياغالي الاثمان",
      "verse": "ياغالي الاثمان غلوك بالحيل",
      "Writer": "خالد الفيصل",
     "Dialect": "Najdi"
   },
   {
      "id": 2,
      "Title": "ياغالي الاثمان",
      "verse": "من يوم سمو بك جميل المحيا",
      "Writer": "خالد الفيصل",
     "Dialect": "Najdi"
   },
   {
      "id": 3,
      "Title": "ياغالي الاثمان",
      "verse": "شخص تمنيته على العدل والميل",
      "Writer": "خالد الفيصل",
     "Dialect": "Najdi"
   }
]
```

Figure 1: Sample of SADSLyC JSON for Najdi Dialect



```
[
{
  "id": 1,
  "Title": "جاني الاسمر",
  "Lyrics": "جاني الاسمر جاني",
  "Writer": "ثريا قابل",
 "Dialect": "Hijazi"
  },
 {
 "id": 2,
  "Title": "جاني الاسمر",
  "Lyrics": "وافتكر زماني",
  "Writer": "ثريا قابل",
 "Dialect": "Hijazi"
  },
  {
 "id": 3,
  "Title": "جاني الاسمر",
  "Lyrics": "وبعد ما نساني جاني",
  "Writer": "ثريا قابل",
 "Dialect": "Hijazi"
 }
  ]
```

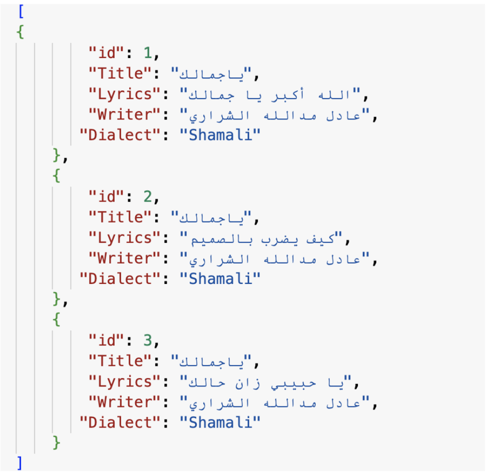
Figure 2: Sample of SADSLyC JSON for Hijazi Dialect

5 Experiments and Results

5.1 Experiments

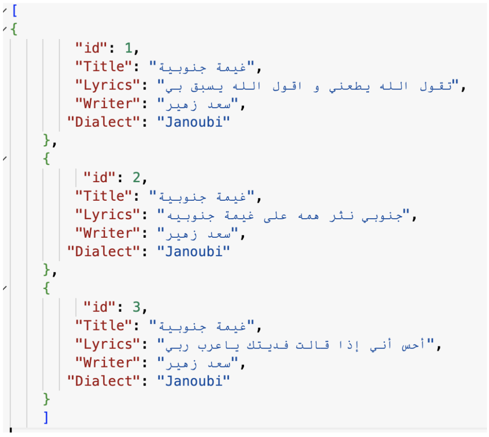
As a baseline experiment, we applied the SaudiBERT model Qarah (2024) for Saudi Arabian dialect identification using the SADSLyC corpus. To address the class imbalance in the SADSLyC corpus, as shown in Table 1, where the Najdi dialect is the dominant class and the other dialects (Hijazi, Janoubi, Shamali, and Shargawi) are underrepresented, we employed a combination of oversampling and stratified splitting. Oversampling was applied during the training phase, specifically increasing the representation of the minority dialects (Shargawi, Shamali, and Janoubi) to create a more

[14] https://github.com/SalwaAlahmari/SADSLyC_Corpus



```
[
{
      "id": 1,
      "Title": "ياجمالك",
      "Lyrics": "الله أكبر يا جمالك",
      "Writer": "عادل مدالله الشراري",
     "Dialect": "Shamali"
  },
  {
      "id": 2,
      "Title": "ياجمالك",
      "Lyrics": "كيف يضرب بالصميم",
      "Writer": "عادل مدالله الشراري",
     "Dialect": "Shamali"
  },
  {
      "id": 3,
      "Title": "ياجمالك",
      "Lyrics": "يا حبيبي زان حالك",
      "Writer": "عادل مدالله الشراري",
     "Dialect": "Shamali"
  }
]
```

Figure 3: Sample of SADSLyC JSON for Shamali Dialect

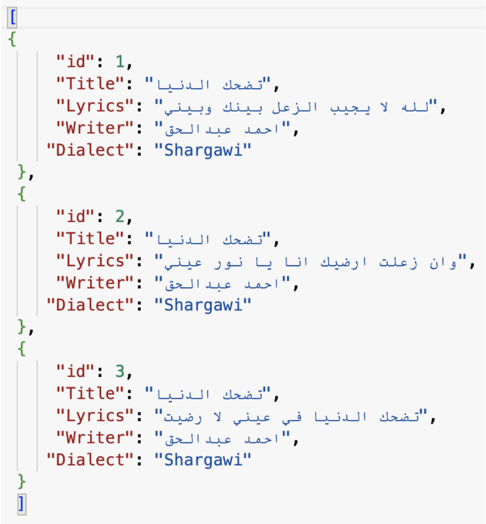


```
[
{
     "id": 1,
     "Title": "غيمة جنوبية",
     "Lyrics": "تقول الله يطعني و اقول الله يسبق بي",
     "Writer": "سعد زهير",
    "Dialect": "Janoubi"
  },
  {
      "id": 2,
     "Title": "غيمة جنوبية",
     "Lyrics": "جنوبي نثر همه على غيمة جنوبيه",
     "Writer": "سعد زهير",
    "Dialect": "Janoubi"
  },
  {
      "id": 3,
     "Title": "غيمة جنوبية",
     "Lyrics": "أحس أني إذا قالت فديتك ياعرب ربي",
     "Writer": "سعد زهير",
    "Dialect": "Janoubi"
  }
  ]
```

Figure 4: Sample of SADSLyC JSON for Jaboubi Dialect

balanced dataset. This adjustment ensures that the model receives adequate samples from each dialect, thereby enhancing its ability to learn distinguishing features from these minority classes and reducing potential bias toward the majority Najdi class.

Additionally, we used stratified splitting when dividing the dataset into training, validation, and test sets. By stratifying based on dialect labels, we ensured consistent class distribution across these subsets, preserving the original corpus proportions. This stratification guarantees that each class is adequately represented during model evaluation, providing a more reliable measure of model performance across all dialects. Combining oversampling with stratified splitting addresses the challenges of imbalanced data, resulting in a model that is better equipped to generalize across all five dialects.



```
[
{
    "id": 1,
    "Title": "تضحك الدنيا",
    "Lyrics": "لله لا يجيب الزعل بينك وبيني",
    "Writer": "احمد عبدالحق",
   "Dialect": "Shargawi"
 },
 {
    "id": 2,
    "Title": "تضحك الدنيا",
    "Lyrics": "وان زعلت ارضيك انا يا نور عيني",
    "Writer": "احمد عبدالحق",
   "Dialect": "Shargawi"
 },
 {
    "id": 3,
    "Title": "تضحك الدنيا",
    "Lyrics": "تضحك الدنيا في عيني لا رضيت",
    "Writer": "احمد عبدالحق",
   "Dialect": "Shargawi"
 }
 ]
```

Figure 5: Sample of SADSLyC JSON for Shargawi Dialect

Accuracy	**Precision**	**Recall**	**F1**
0.73	0.55	0.51	0.53

Table 2: The testing results of dialect identification using SaudiBERT model

5.2 Results

The results of the SaudiBERT model's performance are shown in Table 2, and the confusion matrix is presented in Figure 6. The model achieved an accuracy of 0.73 and an F1 score of 0.53 on the test dataset. These results indicate moderate performance, with potential for improvement in distinguishing between specific Saudi dialects. The confusion matrix reveals that SaudiBERT performs best on the Najdi dialect, with most Najdi samples correctly classified. However, it struggles to differentiate Najdi from other dialects, especially Hijazi and Shamali. Similarly, dialects like Hijazi and Shamali are frequently misclassified as Najdi, suggesting overlapping linguistic features that SaudiBERT finds challenging to separate. Shargawi was the most difficult dialect for the model to classify correctly, with frequent misclassifications into other categories. This is likely due to a combination of limited training data for this dialect and more subtle linguistic distinctions.

Overall, the findings highlight SaudiBERT's strength in identifying prominent dialects like Najdi, but also emphasize the need for further fine-tuning or additional data to improve its ability to capture the nuanced differences among the finer-grained Saudi dialects.

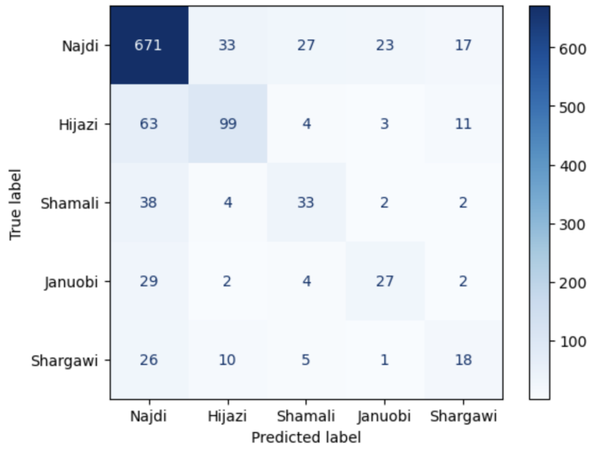


Figure 6: Confusion Matrix of Saudi Arabian dialect identification

6 Discussion

While the study assumes that songwriters from a given region use that region's dialect in their songs, this assumption may not always hold true. For instance, songwriters may prefer to write in Modern Standard Arabic (MSA) rather than their local dialect, such as Shargawi. Furthermore, song lyrics often incorporate multiple dialects as well as MSA. These factors could introduce limitations to the study's assumption, as they may affect the regional representation in the corpus and restrict the findings related to dialect usage.

A deep analysis of a subset of the SADSLyC corpus, based on manual human annotation, reveals dialectal overlap across all dialects, particularly between Najdi and Hijazi. For example, the sentence حنا بحد السيف الدار نحماها, which translates to "We protect our country with the edge of the sword," is labeled as Hijazi in SADSLyC. However, this sentence could be labeled as both Najdi and Hijazi, as it lacks distinctive dialectal features.

7 Conclusion and Future Work

To the best of our knowledge, there is currently no corpus specifically designed for Saudi Arabian song lyrics. The SADSLyC corpus will be the first collection to feature Saudi Arabian songs, representing five major dialects spoken across the country. The experimental results highlight both the strengths and limitations of SaudiBERT for dialect classification, particularly with respect to the fine-grained Saudi dialects, and underscore the need for further fine-tuning on more specialized datasets.

As part of our future research, we plan to expand the SADSLyC corpus by transcribing YouTube videos that showcase a broader range of Saudi songs and dialects.

References

Hamed AL-Rubaiee, Renxi Qiu, Khalid Alomar, and Dayou Li. 2017. Sentiment analysis of arabic tweets in e-learning. *Journal of Computer Science*, 12(11):553–563.

Nora Al-Twairesh, Rawan Al-Matham, Nora Madi, Nada Almugren, Al-Hanouf Al-Aljmi, Shahad Alshalan, Raghad Alshalan, Nafla Alrumayyan, Shams Al-Manea, Sumayah Bawazeer, Nourah Al-Mutlaq, Nada Almanea, Waad Bin Huwaymil, Dalal Alqusair, Reem Alotaibi, Suha Al-Senaydi, and Abeer Alfutamani. 2018. Suar: Towards building a corpus for the saudi dialect. *Procedia Computer Science*, 142:72–82. Arabic Computational Linguistics.

Salwa Alahmari, Eric Atwell, and Mohammad Ammar Alsalka. 2024. Saudi arabic multi-dialects identification in social media texts. In *Intelligent Computing*, pages 209–217, Cham. Springer Nature Switzerland.

Halah AlMazrua, Najla AlHazzani, Amaal AlDawod, Lama AlAwlaqi, Noura AlReshoudi, Hend Al-Khalifa, and Luluh AlDhubayi. 2022. Sa'7r: A saudi dialect irony dataset. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 60–70, Marseille, France. European Language Resources Association.

Latifah Almuqren and Alexandra Ioana Cristea. 2021. Aracust: a saudi telecom tweets corpus for sentiment analysis. *PeerJ Computer Science*, 7.

Adel Assiri, Ahmed Emam, and Hmood Al-Dossari. 2016. Saudi twitter corpus for sentiment analysis. *International Journal of Computer and Information Engineering*, 10(2):272–275.

Afnan Bayazed, Ola Torabah, Redha AlSulami, Dimah Alahmadi, Amal Babour, and Kawther Saeedi. 2020. Sdct: Multi-dialects corpus classification for saudi tweets. *International Journal of Advanced Computer Science and Applications*, 11(11).

Mahmoud El-Haj. 2020. Habibi - a multi dialect multi national Arabic song lyrics corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1318–1326, Marseille, France. European Language Resources Association.

Adam Kilgarriff and Gregory Grefenstette. 2001. Web as corpus. In *Proceedings of the Workshop on Comparing Corpora, 8th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2001)*, Toulouse, France.

Faisal Qarah. 2024. Saudibert: A large language model pretrained on saudi dialect corpora.