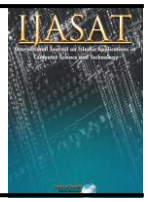

ChatGPT for Identifying Saudi Arabic Dialects

Salwa Saad Alahmari[1,2, a], Eric Atwell[1, b] and Mohammad Ammar Alsalka[1, c]

[1]University of Leeds, Leeds, United Kingdom
[2]University of Hafr Al-Batin, Hafr Al-Batin 39524, Kingdom of Saudi Arabia
[a]SCSSALA, [b]E.S.ATWELL, [c]M.A.ALSALKA@leeds.ac.uk

Abstract
According to Alwakid (2020),Saudipeoplespeak-ingeneral- five main dialects namely; Hejazi (spoken by people in western regions), Najdi (spoken by people in central regions), Sharqawi (spoken by peo- ple in eastern regions), Janubi (spoken by people in southern regions), and Shamali (spoken by people in northern regions) were selected in this study. In this study, we examined the ability of ChatGPT to identify these sub-dialects by collecting a representative sample dataset from Twitter1. The experimental results demonstrate that ChatGPT achieved an overall accuracy of 0.42 in our sample dataset.

Keywords: ChatGPT, Saudi Arabic Dialects, Dialects Identification, Arabic NLP, Social Media.

1. Introduction

In recent years, social networking platforms have experienced significant growth and have become essential hubs of information exchange. These platforms facilitate global communication and sharing of information, with a wealth of data that proves valuable in various Natural Language Processing (NLP) applications like text categorization, sentiment analysis, and machine translation. Users of- ten communicate in different regional dialects of the same language, making it imperative to discern and classify these dialectical variations, a fundamental pro- cess for downstream NLP tasks. This endeavour, known as dialect identification (DI), dialect detection (DD), or dialect recognition (DR), engages experts from linguistic and computer science domains. The ability to identify dialects holds immense importance across diverse fields, including opinion mining, decision- making processes, and marketing strategies.

Arabic, among the oldest Semitic languages, holds a significant place as the fifth most spoken language globally. Beyond its widespread usage, Arabic exhibits a rich array of dialects along with formal and informal variations. Across the Arabic-speaking world, dialectal varieties are categorized into five main groups: Egyptian (EGY), Levantine (LEV), Gulf (GLF), Iraqi (IRQ), and Maghrebi (NOR).

Identifying Arabic dialects poses a formidable challenge due to several factors. Chief among these is the task of distinguishing between closely related dialects with extensive shared vocabularies. Additionally, linguistic code-switching fur- ther complicates this task, as it involves the simultaneous use of multiple Arabic variations within a single sentence. This

[1] https://twitter.com

complexity is heightened in online social media, where users blend Modern Standard Arabic (MSA) with dialectal Arabic, making pinpointing code-switching instances particularly challenging. Moreover, the Arabic script itself presents obstacles, as it may obscure the true sounds and pronunciations of letters and words across different dialects. Furthermore, the emergence of Arabizi, a form of writing Arabic using Latin letters and numbers, adds another layer of difficulty. With no established guidelines or standards for Arabizi usage, identifying Arabic dialects from written texts becomes even more daunting.

Many studies oversimplify by equating the Saudi Arabic dialect with the Gulf dialect, overlooking the diverse linguistic landscape of Saudi Arabia. The country encompasses various regions, each with its distinct dialect. For instance, the eastern region, bordering the Gulf countries, shares a dialectal affinity with the Gulf dialect. Conversely, the western region, notably the Hijazi dialect, exhibits similarities with the Egyptian dialect as opposed to the Gulf. The central region predominantly speaks the Najdi dialect, while the southern and north- ern regions have their unique southern and northern dialects, respectively. This diversity underscores the importance of comprehending the nuances of Saudi Arabic dialects for applications utilizing social media content. Unlike Egyptian or Levantine dialects, existing datasets and lexicons for Arabic dialects are inadequate for Saudi Arabic. Moreover, the scarcity of freely available standardized corpora and lexicons, such as the Golden Standard Corpus (GSC), further com- pounds the challenge. Existing resources for Saudi Arabic dialects are typically inaccessible to the public, requiring prior permission for data and resource reuse.

2. Related Work

A recent study by Almuqren & Cristea (2021) created a gold standard corpus called AraCust consisting of 20,000 Saudi Arabic tweets in the telecom field. They included three Saudi Arabic telecom companies: Saudi Arabic Telcom Company (STC), Mobily, and Zain for Arabic Sentiment Analysis (ASA). Another study byAlMazrua et al. (2022) constructed Sa7'r, an Arabic irony detection corpus for the Saudi Arabic dialects. The tweets were extracted using hashtags, key phrases, and terms regarding irony in Saudi Arabic dialects. The corpus is composed of 19,804 Saudi Arabic tweets classified as irony and non- irony.

In additionAzmi & Alzanin (2014) published a corpus by collecting comments from readers on articles in two Saudi Arabic newspapers, namely Alriyadh and Aljazirah. Their corpus included 815 texts assigned to four groups ranging from strongly positive to strongly negative.Al-Rubaiee et al. (2016)collected 1,331 tweets representing Saudi Arabic dialect about the stock market analysis pro- gram Mubasher. The tweets were categorized as favorable, negative, or neutral with the assistance of two Mubasher employees. Assiri et al. (2016) used Saudi Arabic hashtags to create a corpus of 4,700 tweets. Two annotators classified each tweet as favourable, negative, or neutral according to six specified instruc- tions. The overall observed agreement was .88, and the kappa coefficient was 0.807 based on the location of the tweets. Moreover, Al-Twairesh et al. (2017) compiled a corpus of Saudi Arabic tweets made up of 17,573 tweets taken from more than 2.2 million Arabic tweets from Saudi Arabia. The dataset was categorized by three annotators into four: posi- tive, negative, neutral, and mixed.

3. Corpus Design Criteria

In this research, we aim to create a balanced annotated corpus representing the language varieties used in Saudi Arabia. Moreover, we aim to make this corpus available to be used in different academic research in dialectology and stylistic fields. To achieve these aims, we considered the following design criteria, which were mainly inspired by two sources. The first

is the International Corpus of English (ICE)[2]. And the second is, Arabic Contemporary Corpus (ACC)[3]:

- **The text mode and language**: The mode of the included texts is written machine-readable text. In addition, this corpus represents different Arabic varieties used in the country of Saudi Arabia. Mainly the dialects are spoken in the five main regions of Saudi Arabia.
- **Genre:** The scope of this proposed corpus involves all main Saudi Arabic dialects. This is what makes it distinct from other Arabic corpora that currently exist. The proposed corpus population will include various genres to reflect different linguistic features in Saudi Arabic dialects. These genres are users' posts and comments on Twitter.
- **Time Span:** We covered the period from January 2023 to April 2023.
- **Sampling:** The proposed corpus aims to represent different varieties of Arabic spoken in the country of Saudi Arabia. To achieve this, texts are selected from different regions of Saudi Arabia to represent the main Saudi Arabic dialects. Therefore, the corpus includes each collected text as it is and not a sample of it to preserve all linguistic features in all texts.
- **Gender:** our corpus includes users/writers from both genders. Hence, this enables comparison between the linguistic and stylistic characteristics of language variety used by males and females from different age groups above 18.
- **Anonymization:** we removed any personal information about the online users from the extracted texts to protect their privacy.
- **Balance:** our main objective in this research is to build a balanced corpus for Saudi Arabic dialects. To achieve this, the full extracted text of each genre is included.
- **Representativeness:** to achieve representativeness in the constructed corpus, we have built and used the pre-define lists of dialectal words from the five dialects. Moreover, both genders from various age groups are included.

4. Corpus Construction

The main methodological steps of the corpus construction are shown in Figure 1 and detailed below:

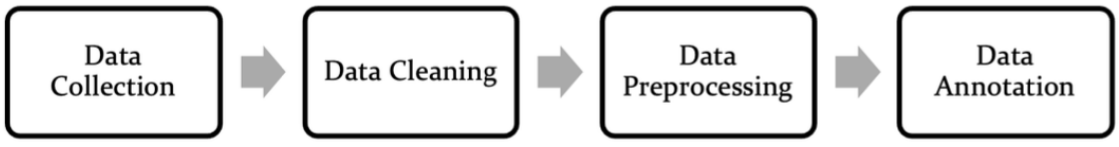

Figure 1. The methodological steps of the corpus construction

4.1 Data Collection

This study aims to assess the efficacy of the fine-tuned ChatGPT model in dis- cerning various Saudi dialects. Encompassing Najdi, Hejazi, Janubi, Shamali and Sharqawi dialects, the

[2] https://www.ice-corpora.uzh.ch/en/design.html
[3] https://eprints.whiterose.ac.uk/82289/1/TheDesignOfACorpusContemporaryArabic2003.pdf

research set out to collect tweets corresponding to trending hashtags within each dialect's respective time zone. However, initial analysis revealed that a significant portion of tweets under these hashtags were unrelated to the specified dialects. Consequently, we opted to pursue a different approach, searching for unique words or phrases distinctive to each dialect. This strategy was intended to yield more relevant data, thus improving the accuracy of dialect identification results. Saudi tweets were acquired utilizing the Twitter API, employing a search method based on predefined lists of keywords specific to each dialect. These lists, comprising the most frequently used words and phrases, were manually compiled with the assistance of native speakers for each dialect. In total, 50 dialectal keywords, with 10 keywords allocated per dialect, were incorporated into this study. Table 1 illustrates examples of the selected dialectal keywords present in Saudi Arabic tweets, along with their corresponding English translations. The collected tweets were then stored in a comma-separated values (CSV) format file, facilitating easy visualization of tweets alongside their annotations in spreadsheet software such as Microsoft Excel.

Table 1. Keywords used for data collection process.

Dialect	Keyword	Example	English Translation
Norther	وش نوحك	انت وش نوحك داخل المنشن حقنا هذا اسرار قبيله شنبك ده حنلئولك	What is your problem enter our mention these are the trip secrets your mustache I will shave it
Janubi	هبولنا	هبولنا كلمات جنوبية محد يعرفها غيرنا يالجنوبيين	Give us words from Janubi accent that are no one can understand them except us
Sharqawi	خلف تشبدي	خلف تشبدي انتين لان عندش جامعه باجر بس اختاري الوقت اللي يناسبش وانتين معزووومه اكيد	After my liver because you have university tomorrow but choice any time suitable for you and you will be invited for sure
Hejazi	الهرجه	ياخوان خلصت التذاكر ايش الهرجه	Brothers, tickets are finished, what is the matter
Najdi	تهقى	تهقى بنتخطى كل شعور حلو عشناه؟	Do you think we can get over every sweet feeling we experienced

4.2 Data Cleaning

In this research, cleaning was done as the second step by removing any Uniform Resource Locator (URL), user mentions, emojis, and any incomplete or ineffective texts, moreover, by removing double spaces, HTML tags, and non-Arabic letters from the text.

4.3 Data Pre-processing

One of the fundamental steps in any NLP task is the pre-processing stage, which involves applying different functions to the given data. As a third step, we automatically applied NLP pre-processing techniques such as normalization, tokenization, and discretization (Tashkeel) removal for Arabic text using CAMeL tools (Obeid et al., 2020). Table 2 shows example of tweets after and before cleaning and pre-processing.

Table 2. Examples of tweets after and before cleaning and pre-processing

Techniques	Tweet before	Tweet after
Remove Emojis	:) خل بقعا تصوعك وتلوعك اقدر أحلف إنك من رجال الجنوب	خل بقعا تصوعك وتلوعك اقدر أحلف إنك من رجال الجنوب
Remove user mention,	@7blaan_ خلني في وجهك ان ضاقت الدنيا عليك كل همٍ لا تضايقت ياخـوي ازهلـه	خلني في وجهك ان ضاقت الدنيا عليك كل همٍ لا تضايقت ياخـوي ازهلـه
Normalize all Alef variations	أمشي وهمي جبل عمشي معي وين ما أروح أزريت أشيل الجبل وأزريت لا أفارقه	امشي وهمي جبل عمشي معي وين ما اروح ازريت اشيل الجبل وازريت لاافارقه
Remove diacritics	عزُي لصدرٍ عاقبته الهواجيس تقول تقطع بالسكاكين جوفه أقوم وأقعد وأتعوذ من ابليس	عزي لصدر عاقبته الهواجيس تقول تقطع بالسكاكين جوفه أقوم وأقعد وأتعوذ من ابليس

4.4 Data Annotation

one of the key challenges in corpus construction is the annotation step, where the corpus compilers must think about the accuracy of text tagging. For our sample data set from Twitter, we have auto-labelled the text with five labels: Najdi, Hejazi, Janubi, Shimali and Sharqawi. The auto-labelling process is done by searching for the existence of any dialectal word from the pre-defined lists in the text. For example, if they found dialectal word is from the Janubi list, then the text will be given a Janubi label.

5. Corpus Validation

According to Tseng et al. (2020), to validate the accuracy of the annotated corpus, there are two hypotheses about corpus validation: The first is that the performance of a machine learning model trained using manually labelled samples should surpass that of the same model trained using randomly re-labelled samples, indicating the importance of accurate labelling. The second is that, in general cases, the classifiers which give the best results with the most available corpora should also give the best results for the corpus under the validation process. This ensures consistency with performance metrics. This research aims to validate our proposed corpus using the above-mentioned hypothesis.

6. Corpus Documentation

our proposed corpus will be produced with detailed documentation which provides information about the sampling methods, text sources, annotation schemes, and other relevant metadata. This documentation allows researchers to under- stand the corpus's construction and use it effectively.

7. Experiments and Results

From the ChatGPT API documentation[4] regarding ChatGPT fine-tuning: we need to create a training dataset to fine-tune ChatGPT to meet certain use cases. This dataset shows examples of the structure and features of the input and output data we would like ChatGPT to learn from. In our case, we built the training data for the Saudi Arabic dialects identification task by

[4] https://platform.openai.com/docs/introduction/overview

providing Saudi Arabic tweets and their classification labels to the predefined ChatGPT model "ada". In data preparation for fine-tuning, we followed the instructions of OpenAI in related to data quality, quantity: and format. Quality means that data is precise and appropriate to the case on hand, the performance of the fine-tuned model depends heavily on the quality level of the training data. The quantity of data is another important standard to fine-tune the predefined deep learning model which we are understanding and putting on consideration for future improvement. JSON (JavaScript Object Notation) is a recommended for- mat for Chat-GPT3 thus we stored our dataset as shown in Fig. 2. To fine tune ChatGPT model we have requested OpenAI API key to access ChatGPT models and to be used in the python code. After the data preparation step and prior to the fine-tuning process, the dataset was divided by ChatGPT into a training set and a validation set. When the fine-tuning process is done, we used the fine-tuned model to predict the dialect of unseen data (in our case Saudi Arabic text) and get the prediction result. Table 3 shows the classification accuracy and weighted F1-score for our data using the fine-tuned model.

prompt	completion
انت وش نوحك داخل المنشن حقنا هذا اسرار قبيله شنبك دهـ حتلئولك	Shamali
نشهد بالعافيه بس هب لى منه حبتين	Janubi
وتظن انك نجوت وتهزمك (اها عاد)	Sharqawi
ايش دا كل الناس اتغيروا حتى السواق!	Hejazi
نفس الشخص اللي داق الثقل عليك قاعد يتهنوص مع غيرك	Najdi

Figure 2. Snippet of dataset file in form of prompt and completion

Table 3. Classification Accuracy and weighted F1-score of the fine-tuned model

Accuracy	F1-Score
0.42	0.41

8. Conclusion

This study aimed to assess the performance of the ChatGPT model following fine-tuning with our sample dataset for Saudi Arabic dialects. We constructed the corpus using lists of distinct keywords for each of the five Saudi Arabic dialects. Manual data annotation involved assigning one of five labels—Najdi, Hejazi, Janubi, Shamali, and Sharqawi—to the collected data. The results in- dicated that the fine-tuned ChatGPT model achieved acceptable accuracy con- sidering the dataset size. This paper presents initial experiments in fine-tuning ChatGPT's predefined models for identifying Saudi Arabic dialects. Addition- ally, this research marks the initial step towards contributing to the development of dictionaries for the less prevalent Shamali, Janubi, and Sharqawi dialects com- pared to the more dominant Hejazi and Najdi dialects in Saudi Arabia.

Due to limitations regarding the size of freely available data for fine-tuning the ChatGPT model, the identification results were lower than expected. For future work, we plan to expand the dataset by incorporating more content not only from Twitter but also from YouTube and Instagram. To augment our Twit- ter sub-corpus, we intend to consider the user's geographical location during the annotation process. Geographical information can be explicitly stated in the user profile or determined using spatial coordinate points for different cities in Saudi Arabia. However, a significant drawback of this method is the potential for mislabeling tweets. For example, all tweets from Riyadh may be labeled as Najdi even if they lack any Najdi linguistic features. To address this drawback, we propose a combined method involving two sub-tasks:

First, identifying the tweet's location, and second, detecting the presence of dialectal words within the tweet.

References

AlMazrua, H., AlHazzani, N., AlDawod, A., AlAwlaqi, L., AlReshoudi, N., Al-Khalifa, H., & AlDhubayi, L. (2022). Sa '7r: A saudi dialect irony dataset. InProceedinsg of the 5th workshop on open-source arabic corpora and processingtools with shared tasks on qur'an qa and fine-grained hate speech detection (pp.60–70).

Almuqren, L., & Cristea, A. (2021). AraCust: a Saudi Telecom Tweetscorpus for sentiment analysis. PeerJ Computer Science, 7 , e510. doi:https://doi.org/10.7717/peerj-cs.510

Al-Rubaiee, H., Qiu, R., & Li, D. (2016). Identifying mubasher software products through sentiment analysis of arabic tweets. In 2016 international conferenceon industrial informatics and computer systems (ciics) (pp. 1–6).

Al-Twairesh, N., Al-Khalifa, H., Al-Salman, A., & Al-Ohali, Y. (2017). Arasenti-tweet: A corpus for arabic sentiment analysis of saudi tweets. Procedia Com-puter Science, 117 , 63–72.

Alwakid, G. (2020). Sentiment analysis of dialectical arabic social media content using a hybrid linguistic-machine learning approach. Nottingham TrentUniversity (United Kingdom).

Assiri, A., Emam, A., & Al-Dossari, H. (2016). Saudi twitter corpus for sentiment analysis. International Journal of Computer and Information Engineering,10 (2), 272–275.

Azmi, A. M., & Alzanin, S. M. (2014). Aara'–a system for mining the polarity of saudi public opinion through e-newspaper comments. Journal of Information Science, 40 (3), 398–410.

Obeid, O., Zalmout, N., Khalifa, S., Taji, D., Oudah, M., Alhafni, B.,Habash, N. (2020, May). CAMeL tools: An open source python. toolkit for Arabic natural language processing. In Proceedings of the 12th language resources and evaluation conference (pp. 7022–7032). Marseille, France: European Language Resources Association. Retrieved fromhttps://www.aclweb.org/anthology/2020.lrec-1.868

Tseng, Y.-H., Wu, W.-S., Chang, C.-Y., Chen, H.-C., & Hsu, W.-L. (2020, May).Development and validation of a corpus for machine humor comprehension.In N. Calzolari et al. (Eds.), Proceedings of the twelfth language resources and evaluation conference (pp. 1346–1352). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2020.lrec-1.168