Saudi Arabic Multi-dialects Identification in Social Media Texts

Salwa Alahmari[1,2][0009−0002−6490−3295], Eric Atwell[1][0000−0001−9395−3764], and Mohammad Ammar Alsalka[1][0000−0003−3335−1918]

[1] University of Leeds, United Kingdome, Leeds
[2] University of Hafr Albatain, Saudi Arabic, Hafr Albatain
{Scssala,E.S.Atwell,M.A.Alsalka}@leeds.ac.uk

Abstract. ChatGPT is a state-of-the-art, robust artificial intelligence language model that can be used in a wide range of Natural Language Processing (NLP) applications. These applications include, but are not restricted to, text classification, text generation, sentiment analysis, and question answering. ChatGPT is primarily aimed at generating English text, but it has also been found to process other languages, such as Arabic language. This paper shows the usage of the ChatGPT model in the task of dialect identification, specifically for Saudi Arabic dialects. Five different Saudi Arabic dialects, namely Hijazi (spoken by people in western regions), Najdi (spoken by people in central regions), Eastern (spoken by people in eastern regions), Southern (spoken by people in southern regions), and Northern (spoken by people in northern regions) were selected in this study. The experimental results demonstrate that ChatGPT achieved an overall accuracy of 0.42, which is higher than identification with a Support Vector Machine (SVM) that gave 0.33 in our sample data-set.

Keywords: ChatGPT · Saudi Arabic Dialects · Dialects Identification· Arabic NLP · Social Media

Introduction

The social networking sites have expanded significantly over the past few years and have emerged as vital informational resources. These networks offer a platform for global communication and information sharing. A significant amount of the data available in these social networks is useful in various NLP applications such as text classification, sentiment analysis, and machine translation. People can speak and write in different dialects of the same languages based on their geographical location. Identifying and distinguishing the different dialects within the same language is a fundamental step for any downstream NLP tasks due to the different features of different dialects. The process of determining the dialect of a given text or speech is known as dialect identification (DI), dialect detection (DD), or dialect recognition (DR). Experts from the fields of linguistic and computer science are involved in the task of identifying dialectal text or speech. The

DI task is crucial for different applications such as opinion mining, decision making, and for marketing purposes. The Arabic language is one of the oldest spoken languages from the group of Semitic languages. It is widely used as it is ranked the fifth most spoken language[1]. In addition to dialects, Arabic has different formal and informal variations (Habash, 2010).The Dialectal Varieties in the Arabic world are mainly divided into five groups namely: Egyptian (EGY), Levantine (LEV), Gulf (GLF), Iraqi (IRQ), and Maghrebi (NOR) (Habash, 2010). Arabic dialect identification (ADI) considers a challenging task for different reasons, the most important one is distinguishing between two Arabic dialects that have a large vocabulary in common and are closely related to each other (Jauhiainen, Lui, Zampieri, Baldwin, & Lindén, 2019). Linguistic Code-switching (LCS) is another reason for the difficulty of the dialect identification task, where a mixture of Arabic variations is used together in the same sentence (Jauhiainen et al., 2019).In online social media, users write text that is a mix of Modern Standard Arabic (MSA) and one or more of dialectal Arabic, which makes the task of identifying the code-switching point a difficult task(Elfardy & Diab, 2013). In addition, some causes come from the Arabic script of the written texts of various dialects. The Arabic script may hide the true sounds and pronunciations of letters and words (Holes, 1984).Moreover, in online conversation, users sometimes feel more comfortable writing Arabic in different forms like Arabizi. Arabizi is a technique for writing Arabic using Latin letters and numbers to represent letters that are absent in the English language. Due to the lack of guidelines or standards for writing in Arabizi, this dialect makes it difficult to identify Arabic dialects from written texts (Jamaly, Darwish, Ahmed, & Hasan, 2014). Most of the research assumes the Saudi Arabic dialect and the Gulf dialect to be the same, which is an oversimplification. Saudi Arabia is a large country with different dialects for different regions. For example, the eastern region, which is the nearest part of Saudi Arabia to the Gulf countries, speaks a dialect that is relatedly close to the Gulf dialect. In the western region, Saudi Arabic people speak Hijazi, which is slightly close to the Egyptian dialect to the Gulf dialect (Al-Twairesh et al., 2018).In the central region of Saudi Arabia, people speak the Najdi dialect. In addition, Saudi Arabic people who live in the south and the north are speaking southern and northern dialects, respectively. As a result, there is a need to understand the unique features of Saudi Arabic dialects in the task of Saudi Arabic Dialects identification (SDI) in order to be used in different applications that use social media content. The available datasets and lexicon for Arabic dialects such as Egyptian or Levantine cannot be applied to Saudi Arabic dialects. Moreover, there are not enough DA datasets and lexicons, particularly the free Golden standard Corpus (GSC) for all Saudi Arabic dialects. (Assiri, Emam, & Al-Dossari, 2018).The few existing resources for Saudi Arabic dialects are not publicly available, and permission must be requested in advance to reuse the collected data and resources (Almuqren & Cristea, 2021). This paper is organized as follows: In Section 2, we review the related previous work in the fields of ADI and SDI. Section 3 gives an in-depth explanation of data collection, and

[1] https://rubric.com/en-US/most-spoken-languages-in-the-world/

Section 4 shows the pre-processing and annotation step. We, then describe data preparation for ChatGPT in Section 5, and Section 6 shows the experimental results. Finally, Section 7 explains the conclusion and future work.

Related Work

The problem of dialect identification has received significant attention in recent years due to the increasing use of dialectal variations in social media and the need for accurate dialect identification systems. In this section, we review related research in ADI and divide them into two parts: all Arabic dialects in general and Saudi Arabic dialects. There are several Saudi Arabic datasets, but none are annotated for dialect identification. Most previous studies on the Saudi Arabic dialect focused on sentiment or emotion analysis. In 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 by(AlMazrua 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 addition(Azmi & 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, Qiu, & Li, 2016)collected 1,331 tweets represent Saudi Arabic dialect about the stock market analysis program Mubasher. The tweets were categorized as favorable, negative, or neutral with the assistance of two Mubasher employees. (Assiri, Emam, & Al-Dossari, 2016)used Saudi Arabic hashtags to create a corpus of 4,700 tweets. Two annotators classified each tweet as favorable, negative, or neutral according to six specified instructions. The overall observed agreement was .88, and the kappa coefficient was 0.807 based on the location of the tweets. Moreover,(Al-Twairesh, Al-Khalifa, Al-Salman, & Al-Ohali, 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: positive, negative, neutral, and mixed. Additionally, there are several other Arabic datasets annotated for dialect, but none are specifically classified for different Saudi dialects.For instance,(Alshutayri & Atwell, 2018b)(Alshutayri & Atwell, 2018a) built a multi-dialectal corpus called Social Media Arabic Dialect Corpus (SMADC) by extracting dialectal text from Twitter, Facebook, and online newspapers. This study included the main Arabic dialects: Gulf Dialect, Iraqi Dialect, Levantine Dialect, Egyptian Dialect, and North African Dialect. Furthermore,(Mubarak, 2018) collected Arabic tweets in different Arabic dialects and converted them into Modern Standard Arabic (MSA).The translation was done by hiring native speakers from CrowdFlower[2].

[2] https://visit.crowdflower.com

The four selected dialects in this corpus are: Egyptian, Maghrebi, Levantine, and Gulf. Additionally, the Multi Arabic Dialect Applications and Resources corpus (MADAR)(Bouamor et al., 2018) is a large parallel text collection of 25 city dialects from 15 Arab countries. It includes sentences from the Basic Travel Expression Corpus (BTEC) translated into five regional dialects of Arabic: Maghrebi, Nile Basin, Levant, Gulf, and Yemen. Another corpus was published by(Abdul-Mageed, Zhang, Bouamor, & Habash, 2020) in a shared task called Nuanced Arabic Dialect Identification shared task dataset (NADI). NADI is a collection of naturally produced dialectal Arabic tweets at the sub-country level, identified by province, sent over 10 months. Likewise,(Abdelali, Mubarak, Samih, Hassan, & Darwish, 2021)built the Qatar Computing Research Institute (QCRI) Arabic Dialects Identification (QADI) corpus. QADI is a dataset for identifying Arabic dialects at the country level. The tweets were selected from 18 countries in the Arabic-speaking world and organized in dataset files by country name. There is a growing interest in analyzing the content of Saudi Arabic people on social media networks to assess customer satisfaction in different business markets. However, only a few research projects focus on Saudi Arabic dialect identification and there is no public corpus that includes all Saudi Arabic dialects. Thus, there is a great need for a comprehensive and publicly available corpus for the Saudi Arabic dialect identification challenge.

Data Collection

The objective of this study is to evaluate the performance of the fine-tuned ChatGPT model in identifying Saudi dialects. This research covers all dialects in Saudi Arabia including Northern, Southern, Eastern, Hijazi and Nadji. In the beginning of the data collection stage, we intended to collect tweets from trending hashtags in the time zone for each Saudi dialect. Unfortunately, our analysis of the tweets in these hashtags showed that the majority of the tweets were not related to the dialect in the specific time zone. As a result, we decided that searching for words or phrases which are special from each dialect will help in gathering more related data and hence give better identification results. Saudi tweets were collected using Twitter API by searching for word or phrase in the pre-defined list of keywords for each dialect. The lists of keywords of the most common used word and phrases were built manually with a help of native speakers for each dialect. The total number of dialectal keywords for all Saudi dialects that are used in this study is 50,for each dialect there are 10 dialectal keywords. Table 1 shows examples of the selected dialectal keywords in Saudi Arabic tweets used for each dialect and their English translation. Collected tweets were saved in a comma separated values(CSV) format file, making it easy to display the tweets with their annotation as a spreadsheet in 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 حنلثولك	What is your problem enter our mention these are the trip secrets your mustache I will shave it
Southern	هبولنا	هبولنا كلمات جنوبية محد يعرفها غيرنا يالجنوبيين	Give us words from Southern accent that are no one can understand them except us
Eastern	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
Hijazi	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

Data pre-processing and annotation

One of the fundamental steps in any NLP task is the pre-processing stage, which involves different functions on the given data. Table 2 shows examples of tweets before and after cleaning and pre-processing techniques. In this study, pre-processing was done by first manually cleaning the Saudi Arabic tweets from any Uniform Resource Locator (URL), user mentions, emojis, and any incomplete or ineffective tweets. As a second stage, we automatically applied NLP pre-processing techniques such as normalization, tokenization, and diacritization (Tashkeel) removal for Arabic text using CAMeL tools(Obeid et al., 2020).We then manually annotated the collected data with five labels (Northern, Southern, Eastern, Hijazi, and Najdi). There are two reasons for the manual annotation of the dataset. First, we mainly used the predefined list of keywords for each Saudi Arabic dialect in data collection, so we know the specific dialect. Second, the dataset is small (only 200 tweets).Fig. 1demonstrates the distribution of the dialects in our sample dataset.

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

Techniques	Tweet before	Tweet after
Remove Emojis	:) خل بقما تصوعك وتلوعك اقدر أحلف إنك من رجال الجنوب	خل بقما تصوعك وتلوعك اقدر أحلف إنك من رجال الجنوب
Remove user mention,	@7blaan_ خلني في وجهك ان ضاقت الدنيا عليك كل همٍ لا تضايقت ياخوي ازهله	خلني في وجهك ان ضاقت الدنيا عليك كل همٍ لا تضايقت ياخوي ازهله
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 وأتموذ من ابليس	عزي لصدر عاقبته الهواجيس تقول تقطع بالسكاكين جوفه أقوم وأقعد وأتموذ من ابليس

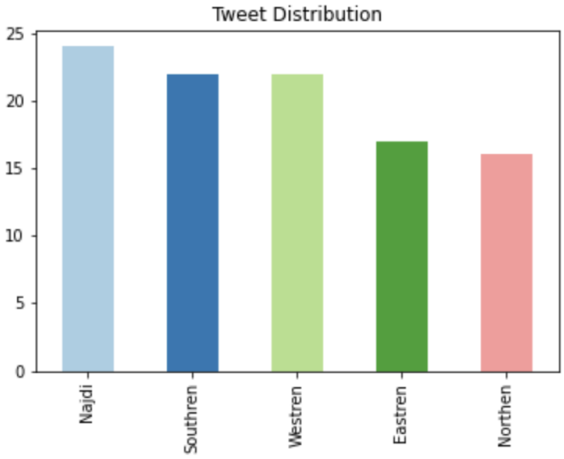

Fig. 1. Dialects distribution on the data-set.

Dataset preparation for ChatGPT

From the ChatGPT API documentation[4] regarding ChatGPT fine tuning: we need to create a training data-set in order to fine-tune ChatGPT to meet certain use cases. This data-set 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 Saudi Arabic dialects identification task by providing Saudi Arabic tweets and their classification labels to predefined ChatGPT model "ada". In data preparation for fine tuning, we followed the instruction 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 predefined deep learning model which we are understanding and put on consideration for future improvement. JSON(JavaScript Object Notation) is a recommended format for Chat-GPT3 as shown in Fig. 2.

Experiments and Results

We applied a baseline experiment using a SVM classifier with RBF kernel to compare the results with fine-tuned ChatGPT model.The data-set was divided by ChatGPT into a training set and validation set. Table 3 shows the classification accuracy and weighted F1-score for our data using the two models. 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 preperation step and

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

Fig. 2. Structure of JSON file format used for fine tuning process.

when the fine tuning process done in 4 epochs as shown in Fig. 3. As a finall step we get the fine-tuned model and we can pass to it any prompt (in our case Saudi Arabic text) and get the prediction as shown in Fig. 4.

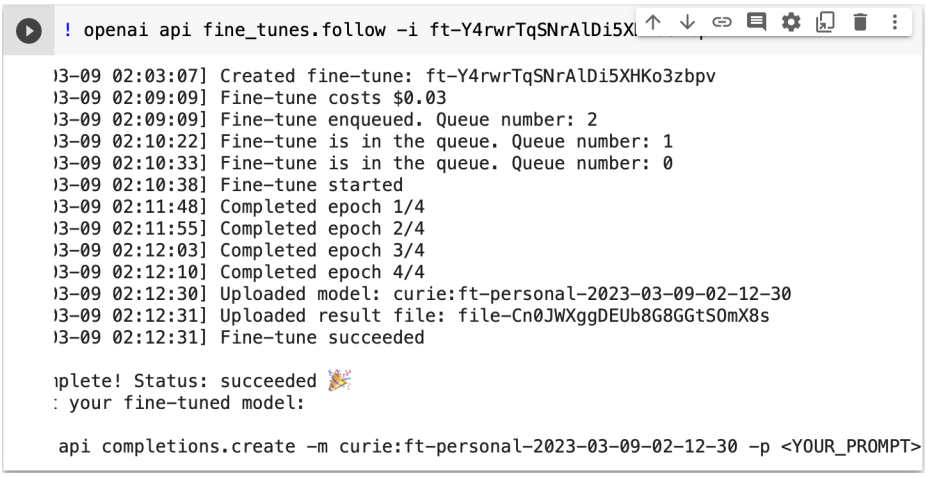

Fig. 3. A snippet of Python code showing fine tuning process.

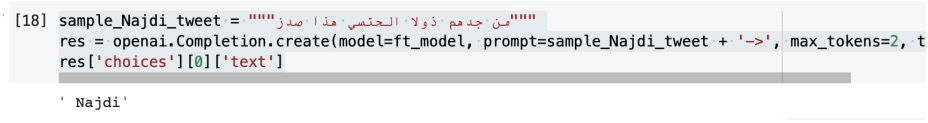

Fig. 4. A snippet of Python code showing using fine-tuned ChatGPT model for prediction.

Table 3. Classification Accuracy and weighted F1-score using the two models

Model	Accuracy	Wieghted F1-Score
SVM Model	0.33	0.33
ChatGPT Model	0.42	0.41

Conclusion

The objective of this paper was to evaluate the performance of ChatGPT model after fine tuning with our sample data-set for Saudi Arabic dialects. The corpus was constructed using lists of different keywords for each the five Saudi Arabic dialects. The data annotation was done manually by adding five labels Northern, Southern, Eastern, Hijazi and Najdi to the collected data. A SVM classfier with RBF kernel was applied on the data-set to compare the classification results with the performance of the fine-tuned ChatGPT model. The results show that the fine-tuned ChatGPT model gave better results than classical classifier with SVM.This paper gives primary experiments in fine tuning ChatGPT predefined models for Saudi Arabic dialects identification. In addition, this research represents a first step to contribute in building dictionaries for Northern, Southern and Eastern dialects dialects which are less popular than Hijazi and Najdi IN Saudi Arabia. For reasons related to the limitation on the free data size to fine tune ChatGPT model the identification results were lower than expected. For future work, we are planning to make the data-set bigger by adding more content from not just Twitter but also YouTube and Instagram.

References

Abdelali, A., Mubarak, H., Samih, Y., Hassan, S., & Darwish, K. (2021, April). QADI: Arabic dialect identification in the wild. In *Proceedings of the sixth arabic natural language processing workshop* (pp. 1–10). Kyiv, Ukraine (Virtual): Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.wanlp-1.1

Abdul-Mageed, M., Zhang, C., Bouamor, H., & Habash, N. (2020). Nadi 2020: The first nuanced arabic dialect identification shared task. In *Proceedings of the fifth arabic natural language processing workshop* (pp. 97–110).

AlMazrua, H., AlHazzani, N., AlDawod, A., AlAwlaqi, L., AlReshoudi, N., Al-Khalifa, H., & AlDhubayi, L. (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* (pp. 60–70).

Almuqren, L., & Cristea, A. (2021). AraCust: a Saudi Telecom Tweets corpus 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 conference on industrial informatics and computer systems (ciics)* (pp. 1–6).

Alshutayri, A., & Atwell, E. (2018a). Arabic dialects annotation using an online game. In *2018 2nd international conference on natural language and speech processing (icnlsp)* (pp. 1–5).

Alshutayri, A., & Atwell, E. (2018b). Creating an arabic dialect text corpus by exploring twitter, facebook, and online newspapers. In *Osact 3 proceedings.*

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

Al-Twairesh, N., Al-Matham, R., Madi, N., Almugren, N., Al-Aljmi, A.-H., Alshalan, S., ... Alfutamani, A. (2018). Suar: Towards building a corpus for the saudi dialect. *Procedia Computer Science*, *142*, 72-82. Retrieved from https://www.sciencedirect.com/science/article/pii/S187705091832163X (Arabic Computational Linguistics) doi: https://doi.org/https://doi.org/10.1016/j.procs.2018.10.462

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.

Assiri, A., Emam, A., & Al-Dossari, H. (2018). Towards enhancement of a lexicon-based approach for saudi dialect sentiment analysis. *Journal of Information Science*, *44*(2), 184-202. Retrieved from https://doi.org/10.1177/0165551516688143 doi: https://doi.org/10.1177/0165551516688143

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.

Bouamor, H., Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Abdulrahim, D., ... others (2018). The madar arabic dialect corpus and lexicon. In *Lrec.*

Elfardy, H., & Diab, M. (2013, August). Sentence level dialect identification in Arabic. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 456–461). Sofia, Bulgaria: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P13-2081

Habash, N. (2010). *Introduction to arabic natural language processing* (1st ed., Vol. 3). Morgan and Claypool Publishers. doi: https://doi.org/10.2200/S00277ED1V01Y201008HLT010

Holes, C. (1984). *Colloquial arabic of the gulf and saudi arabia.* Routledge & Paul.

Jamaly, S., Darwish, N., Ahmed, I., & Hasan, S. (2014). A short review on reverse osmosis pretreatment technologies. *Desalination*, *354*, 30–38.

Jauhiainen, T., Lui, M., Zampieri, M., Baldwin, T., & Lindén, K. (2019). Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, *65*, 675–782.

Mubarak, H. (2018). Dial2msa: A tweets corpus for converting dialectal arabic to modern standard arabic. *OSACT*, *3*, 49.

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 from https://www.aclweb.org/anthology/2020.lrec-1.868