MEETING CHALLENGES OF MODERN STANDARD ARABIC AND SAUDI DIALECT IDENTIFICATION

Yahya Aseri, Khalid Alreemy, Salem Alelyani, Mohamed Mohanna

Center for Artificial Intelligence, King Khalid University, Saudi Arabia

ABSTRACT

Dialect identification is a prior requirement for learning lexical and morphological knowledge a language variation that can be beneficial for natural language processing (NLP) and potential AI downstream tasks. In this paper, we present the first work on sentence-level Modern Standard Arabic (MSA) and Saudi Dialect (SD) identification where we trained and tested three classifiers (Logistic regression, Multi-nominal Na¨ıve Bayes, and Support Vector Machine) on datasets collected from Saudi Twitter and automatically labeled as (MSA) or SD. The model for each configuration was built using two levels of language models, i.e., unigram and bi-gram, as feature sets for training the systems. The model reported high-accuracy performance using 10-fold cross- validations with average 98.98%. This model was evaluated on another unseen, manually-annotated dataset. The best performance of these classifiers was achieved by Multi-nominal Naïve Bayes, reporting 89%.

KEYWORDS

Dialect Identification, NLP, Standard Arabic, Saudi Dialect, Classification.

1. INTRODUCTION

Human language understanding and generation is an essential component for developing numerous AI systems, such as virtual assistants, chatbots, talking robots. This component requires lexical resources, morphological and grammatical knowledge, and meaning representation that captures users' intents and facilitates human-machine interaction, in particular conversational interface, in an efficient and powerful way. Though Arabic natural language processing, as tools for human- machine interaction, has received considerable attention, the differences between spoken/dialectal and standard Arabic pose challenges to natural language processing and potential AI applications [1]. The NLP tools developed for Modern Standard Arabic (MSA) often fail in dealing with mod- ern varieties of Arabic. Dialectal identification (DI), however, is considered an important NLP task that can be utilized for developing lexical resources and prepossessing large-scale data used for machine learning tasks.

Dialectal identification is a form of Language Identification (LI), which is the task of detecting the natural language that a document or part thereof is written in so that a system can mimic the human ability of recognizing certain languages [2]. Researchers in this area do not make a distinction between languages and language varieties/dialects since the computational methods used are identical and challenges faced are similar. Furthermore, the motivation for LI or DI is also almost the same. Though LI was initially motivated by machine translation, it is considered a fundamental component for natural language processing of languages with a high degree of dialectal variety. To ensure that a given document is relevant to NLP tools available, LI is used to

David C. Wyld et al. (Eds): EMSA, SEA, AIFU, NLCAI, BDML, BIoT, NCOM, CLOUD, CCSEA, SIPRO - 2022
pp. 341-351, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.120628

determine the language of the document and whether it is subject to further natural language processing. Moreover, LI plays a vital role in creating lexical resources and corpora used in machine learning tasks, in particular for low- resource languages or dialects.

A major challenge, for Arabic NLP, comes from the fact that Arabic language exists in a state of diglossia [3] in which the standard form of the language, MSA, and the regional dialects live side-by-side and are closely related [4]. While MSA refers to the language used in Arab world used in education, newspapers, and laws documentation, Dialectal Arabic (DA) refers to spoken language (or informal written language) used in daily communication. Spoken Arabic exhibits several language variations, which are a mixture of MSA and a number of Arabic vernaculars. These language variations are what people in Arab world acquire and speak at home and use in their daily lives. From a natural language processing perspective, there are five major groups of dialects that are regionally defined: Egyptian, Gulf, Iraqi, Levantine, and Maghrebi [1, 5]. These variations differ from one another in terms of lexical, morphological and syntactic structures, though they have the core grammar in common. In addition, each group shows internal variations that cannot be neglected. For example, Gulf dialects include Bahraini, Kuwaiti, Omani, Qatari, Saudi, UAE, and Yamani [5]. As we will see in section 2, the aforementioned linguistic diversity has been taken into consideration by the Arabic NLP community and two machine learning models have been proposed. One is a multi-way classification model where the classes range from 3 to 29 classes. The other model is binary-classification model where the task is to distinguish a certain dialect from MSA.

2. RELATED WORK

Arabic dialects identification has recently attracted the Arabic NLP researchers and practitioners [6, 7]. Studies presented, however, differ in terms of their aims, target dialects, and approaches. Some focus on systems performing binary classification between MSA and a specific dialect [4, 8, 9], others describe multi-way classifications between (MSA) and other dialects, including the five major dialects: Egyptian, Gulf, Iraqi, Levantine, and Maghrebi [9– 18].These studies have implemented different methods of traditional machine learning and deep learning algorithms [2].

For multi-dialectal identification, Harrat et al. [11] describe experiments using datasets representing Maghrebi dialects and what they call Middle Eastern dialects as well as MSA. Shervin Malmasi et al. [14] present work on sentence-level Arabic dialects identification. Using a set of surface character and word features, they trained their system on a multidialectal parallel corpus of Arabic. This work shows 74% accuracy on a 6-way multi-dialect classification. Mohamed Lichouri et al. [12] describe methods for textual Arabic dialects identification. The experiments were conducted on two datasets: one represents Maghrebi and Middle Eastern dialects, while the other represents Algerian dialects. For the Middle Eastern dialects, the system achieved an average accuracy of 92% and 76% for Algerian dialects. Leena Lulu et al [15] describe deep learning models used for the automatic classification of Arabic dialectal texts. They used the Arabic Online Commentary (AOC), which includes Egyptian (EGP), and Gulf (GLF), and Levantine (LEV) dialects. Mohamed Ali [16] introduces systems submitted to the Arabic dialect identification shared task 2018, which included MSA, Egyptian, Gulf, Levantine, and north African dialects. For this task, he used character-level convolution neural network as well as dialect embedding vectors, achieving 57.6% F1-score. Mohamed Elaraby et al. [18] used the AOC for both binary and multi-way classification. Having benchmarked the data, they trained and tested six different deep learning methods and compared the results to several classical machine learning models, showing 87.65% accuracy on the binary task (MSA vs. dialects), 87.4% on the three-way dialect task (Egyptian vs. Gulf vs. Levantine), and 82.45% on the four-way variants task (Egyptian vs. Gulf vs. Levantine vs. MSA).

In an attempt to provide a fine-grained classification, Sadat et al. [10] present work on 18 local Arabic dialects. To develop probabilistic models, they used the character n-gram Markov language model and Naive Bayes classifiers trained on datasets derived from social media. Abdul Mageed et al. [19] also describe work for detecting dialects from 29 cities in 10 Arab countries. Similarly, Mohammad Salameh et al. [20] developed a fine-grained system with 25- way classification where the labels are 25 cities from several countries (including Riyadh and Jeddah in Saudi Arabia) as well as MSA. Their systems were trained to predict the location/city of the speaker rather than to give linguistic labels, i.e., dialectal classes, to a given text For binary classification tasks, Elfardy et al. [4, 8, 9] introduce a system performing binary classification between EGP and MSA. Elfardy et al. [4] present work for sentence-level binary classification task performed on EGP and MSA. They implemented supervised machine learning algorithms to train their system, using token level features and other meta features, to predict the correct label for a given sentence. The system achieved an accuracy of 85.5% on the AOC dataset. Tillmann et. al. [8] present another work to perform the same task, i.e., classification between EGP Arabic and MSA. The system was also tested on the AOC dataset and achieved an accuracy of 89.1 %. However, they indicate that the system's performance decreased when evaluation on data from another source. Al-Badrashiny el Al. [9] also focus on MSA and EGP. However, they describe a hybrid approach in which a sentence-level classifier was trained to predict the correct class for each sentence using labels and the confidence scores generated by two underlying classifiers. Their system achieved an accuracy of 90.8%.

The focus of this paper is on Saudi dialect (SD) identification. To the best of our knowledge, this work presents the first identification system of SD. To avoid the over-fitting or under-fitting problems that may result from the linguistic differences among Arabic dialects, we adopt the binary classification model and introduce a system that perform binary classification between MSA and SD. Such a system takes Saudi Twitter texts as inputs and provides linguistic labels for each as either MSA or SD. Because there is a considerable overlap between Arabic dialects and MSA in terms of lexical, morphological, and syntactic properties, we define SD, in the present study, as any text that contains at least one token that lexically or morphologically belongs to the dictionary of SD defined in section 4.2.

3. RESEARCH PROBLEM AND DATA

In this paper, we present a system that discriminates between SD and MSA, which is to our knowledge the first work aiming at this goal. This is a significant step toward building a large lexicon and NLP tools used for AI systems that can understand this dialect.

Given the fact that dealing with spoken forms of Arabic language is not an easy task, Twitter is considered a good source for achieving this goal for one main reason. Twitter contains informal texts that are to a large extent close to spoken language in terms of the lexicon used in this dialect and morphological variations. However, Twitter also contains linguistic data that represent MSAas well as data with code switching from MSA to SD and vice versa. To make use of Twitter in learning this dialect, it is important to first distinguish what can be pure MSA and what represents SD. Hence, this paper presents a system performing binary classification, which takes a sentence as an input and labels it either MSA or SD. SD, in this paper, is linguistically defined based on distinctive morphological properties (i.e., inflectional features) or lexical items (i.e., functional words) used in spoken language. Consequently, texts classified as SD are those containing code- switching between MSA and SD.

4. EXPERIMENTAL SETUP

We conducted the following experiment using three supervised machine learning algorithms: logistic regression (LR), multi-nominal Naïve Bayes (MNB), and support vector machine (SVM). These models were trained on 346,931 tweets collected from Masfah, an online platform, which has a huge number of tweets documents stored by crawling Twitter social network. The algorithms were trained on two subsets of data representing the two linguistic classes: MSA and SD. To train our model to identify SD and distinguish it from MSA texts, we need large-scale training data that represent these two classes. Because manual annotation is time-consuming and requires too much effort, we prepared our training data by automatic extracting and labeling tweets representing each class. The following sub-sections explain text preprocessing, building dictionaries, automatic labeling, model training and evaluation.

4.1. Text Preprocessing

Twitter posts are actually noisy data. Tweets are intended to contain texts, but also contain a mixture of other data types like images and videos. The cleaning process of the data includes the following tasks:

- Removing images, videos, hashtags, user mentions, symbolic characters, emojis, numeric characters, and hyperlinks.
- Elimination of Arabic diacritics *Tashkeel*.
- Replacing *Alif Hamza* (أ) with *plain Alif* (ا).
- Replacing *Taa Marbootah* (ة) with *Haa* (ه).
- Removing stop words.
- Deleting duplicate documents.

The removal of stopwords needs to be done carefully. Not all stopwords should be eliminated. Stopwords like other tokens can be true identifiers and hence many stopwords belong to what we call identification terms that distinguish the MSA sentences from SD sentences. To only eliminate stopwords list that is useless, we used Count Vectorizer for bag-of-words representation rather than TF-IDF. Moreover, common stopwords that appear in both MSA sentences and SD sentences were eliminated because they have no effect on our model. We also eliminated duplicate documents resulting from the removal of hashtags, user mentions, hyperlinks, etc. Such duplicate documents came out to be a problem since they have no value, which results in a negative effect on model training.

4.2. Building Dictionaries

Prior to building lexicons and preparing our training corpus, we had to make a few linguistic assumptions. First, we assume that MSA and SD represent two language levels that lexically and morphologically overlap. Secondly, each level may have its unique lexicon and grammatical properties. Thirdly, speakers of SD often use code switching from SD and MSA and vice versa. Taking these assumptions into consideration, we started with a short list of vocabulary that uniquely identify SD. This list was identified based on three types of features: lexical features, functional words, and inflectional morphology. Likewise, we came up with another short list of vocabulary that uniquely identify MSA and commonly used in formal writing. The two lists were revised and approved by linguists who are experts in both MSA and SD. Examples are listed in Table 1. The list of Vocabulary in each domain cannot belong to both domains at the same time. That is, each word is considered to belong to either MSA or SD. For example, the word "ﻻ أ" is considered an SD identifier as it is widely used in spoken SD but cannot appear in MSA texts. On

the other hand, "فاهم" is a functional word that belongs to MSA; and thus, it is considered an MSA identifier that cannot be an SD identifier at the same time. We used these two short lists to extract text/sentences and group them into two sets of documents: MSA documents and SD documents. To expand the vocabulary lists that identify each class, we extracted all words from all texts in each document set and added them to the corresponding list, which create two dictionaries. The texts belonging to SD document set undoubtedly contain words that belong to MSA list, according to the third assumption previously stated. Thus, we removed the overlap section by eliminating the intersection between the two dictionaries. This process has resulted in a dictionary of pure MSA that contains 41861 words and another one that is pure SD with 51286 words.

4.3. Automatic Annotation

We used the updated dictionaries to re-filter the two class of documents: MSA documents and SD documents. The outputs are two classes that overlap because the informal texts/dialect contain MSA words. Figure 1 represents such an overlap between the two class of documents.

Documents lying in the overlapping region are either MSA documents that have SD identification words or SD documents that includes MSA identification words. Therefore, we eliminated these documents from our training datasets, which decreases the size of the overlap documents. This process of reducing the overlap size yielded a corpus that contains two datasets representing MSA (107,917 sentences) and SD (128,051 sentences). Sentences/documents of each class were automatically labeled either MSA or SD. The two datasets were combined and shuffled to ensure they were normally distributed and to avoid model overfitting.

Table 1: Examples of Tokens used as identifiers of MSA and SD

Standard Token			Dialect Token		
كيفما	ذهبت	استيقظ	بشويس	اهلين	ابغا
يؤدي	ربما	التي	بعدين	ايش	ابغى
لذلك	ريثما	الذي	بياخذ	بالمره	اركد
لقد	سوف	الذين	بيطلع	بايخ	اروح
لكي	سيقوم	حسبما	تبون	برضه	اشلون
من أجل	عندما	حيث	تبي	برضو	اشوف
منذ	فقد	حيثما	على راسي	بروح	الحين
هكذا	كما	ذهبنا	ترا	بس	اللي

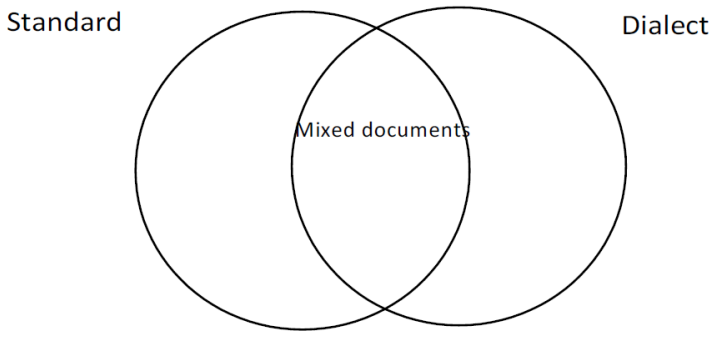


Figure 1: Documents Overlap

4.4. Model Training

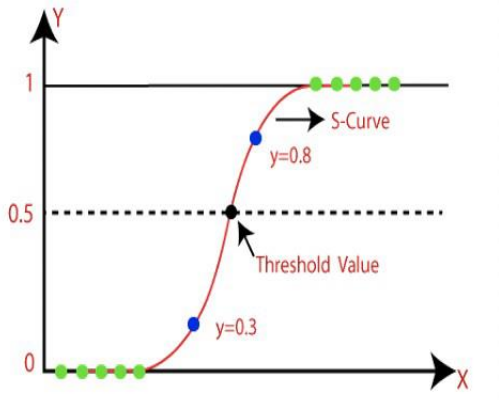
Table 2: Examples of Bi-gram of SD Expression Formed from two MSA Words

Dialect Term	Standard Word-2	Standard Word-1
زي الناس	← الناس	زي
يعطيك العافية	← العافية	يعطيك
على طول	← طول	على

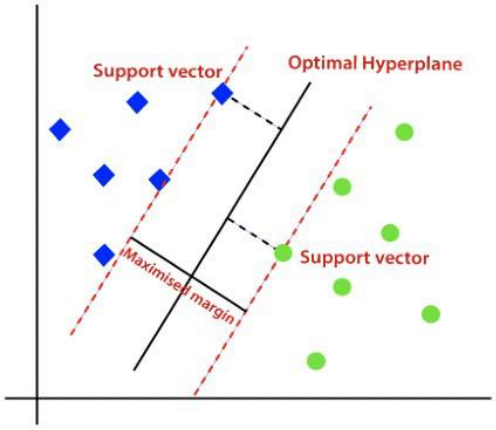
To train our models, each document/sentence was converted into vectors representation using CountVectorizor. The minimum document frequency was set to 10 to emphasize the effectiveness of the selected features. Features were extracted as n-gram terms where n was set to (1,2). We used bi-gram because in many cases an MSA word when combined with another MSA word form an SD phrase/expression. Examples of this phenomenon are shown in Table 2. We implemented three supervised machine learning algorithms to predict discrete values of (0 and 1) as MSA or SD respectively. We used Multi-nominal Naïve Bayes (MNB) classifier, which depends on Bayesian theorem described by the following formula:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Despite its simplicity, it in many cases outperforms other classification algorithms, as we see in our case. Thus, it is widely used in text classification. We also used Logistic Regression (LR). As shown in Figure 2, LR depends on the notion of probability and uses sigmoid function to convert continuous values into discrete numbers, which fits binary classification problems. In our case, LR assigns a probability that a given text belongs to a certain class. Finally, we implemented Support Vector Machine (SVM), an algorithm that can be used for both regression and classification problems. We used SVM because it can handle data with large features, as in our case, in which every data point is plotted in an n dimensional space. Such a classification task is done by finding the hyperplane that separates these spaces. The maximum distance between the nearest data points in every separated spaces is called margin, as shown in Figure 2.



a) Logistic Regression



b) Support Vector Machine

Figure 2: Difference between LR and SVM

Table 3: Accuracy Average of 10-fold Cross-validation

Algorithm	MNB	LR	SVM
Accuracy Average	98.24	99.44	99.28

The classifiers take n-gram of an input sentence and compute the probability/likelihood that this sentence belongs to the MSA class or the SD class. Instead of splitting the data into two subsets, we used k-fold cross-validation where k=10. The dataset was split into 10-folds and each model was repeatedly built using 90% for training while holding 10% for testing. For each model configuration, the accuracy of each fold was captured. Table 3 shows that the three models perform with a slight difference. The 10-fold average score for MNB, LR, and SVM is also reported as 98.24%, 99.44% and 99.28% respectively.

5. MODEL EVALUATION AND DISCUSSION

To evaluate the models' performance, the three classifiers were tested on an unseen, manually doubled-annotated dataset. The point here is to compare the model performance on this dataset to its performance on the dataset that was automatically labeled. The total number of this testset is 15747 Saudi tweets. Unlike training data that were automatically collected and labeled, this dataset was randomly collected from Saudi Twitter and manually annotated by language experts. It was cleaned and preprocessed, using the same tools mentioned earlier, and given to annotators who label each sentence as either MSA or SD. Figure 4 shows the performance at a probability threshold of 0.5. True Positive Rate (TPR) refers to the ratio of correctly predicted positive labels from all the positive labels, which is the MSA in our case, while False Positive Rate (FPR) refers to the ratio of incorrectly predicted positive labels from all the negative labels which is the SD. We computed their values as follows.

$$True\ Positive\ Rate\ (TPR) = \frac{TP}{TP + FN}$$

$$False\ Positive\ Rate\ (FPR) = \frac{FP}{FP + TN}$$

As we had 107,917 records for MSA and 128,051 records for SD in our training data, the percentages of MSA and SD with respect to the total records are 46% and 54% respectively. Such data can be considered balanced data. The predicted values for the validation data are considered either "0" for MSA or "1" for SD. By using the threshold value of 0.5, the prediction is considered as "0" when the predicted probability is in the range [0.0 - 0.49], and "1" when the prediction is in the range [0.5 – 1.0]. In contrast with the models' performance reported above, Table 4 shows the performance these classifiers where the best result was achieved by by MNB classifier, reporting 89.05% for the model accuracy and 88% for F- score.

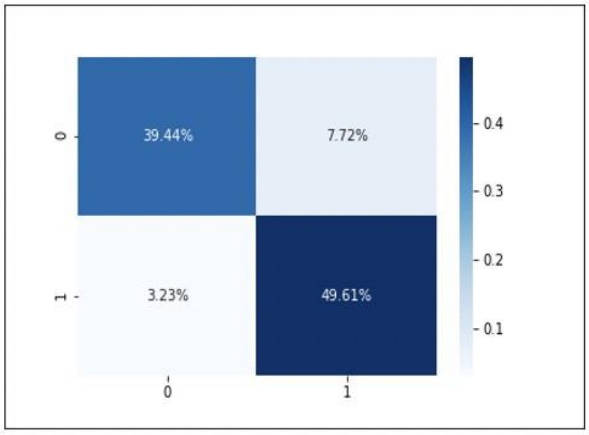


Figure 3: Confusion Matrix

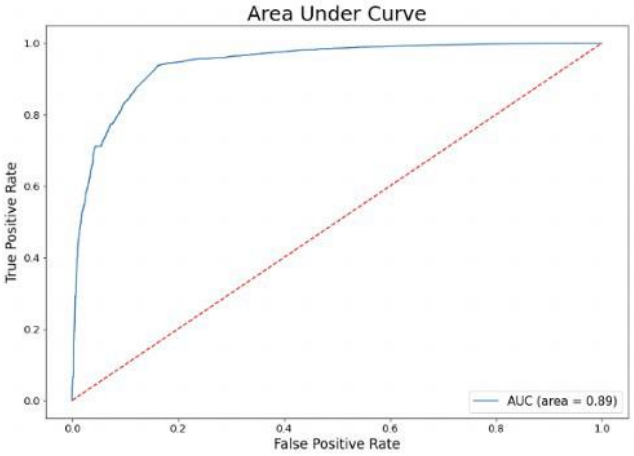


Figure 4: Area Under the Curve

Table 4: Models' Performance on Double-Annotated Data

	Accuracy	**Precision**	**Recall**	**F-score**
Multinomial Naïve Bayes	89.05	0.92	0.84	0.88
Logistic Regression	63.31	0.94	0.24	0.38
Support Vector Machine	63.41	0.90	0.25	0.39

This experiment shows that MNB performs well on SD identification, though it is a difficult task because the high degree of similarity between MSA and SD and studies have shown the difficultyof language/dialect identification task performed on neighboring dialects or similar

languages [2, 11, 12, 14]. The drop in the result may be attributed to the fact that in SD a sentence that does not contain any dialectal token can also be recognized by human annotators as SD when it hasa combination of two or three MSA words that form a SD phrase, as discussed previously. This linguistic phenomenon cannot be recognized by the system unless we have a large lexicon of n-gram(where n > 1) of SD, which is not available yet. In addition, names entities represent a challenge to any language or dialect identification task. In this experiment, these entities have been treatedin our language model as regular tokens. We may also consider text normalization as another issue that need to be taken care of in such a task. We believe that deep and careful text normalization is still required prior to training our model.

6. CONCLUSION AND FUTURE WORK

In this paper, we presented work on sentence-level Saudi Dialect (SD) identification task where we trained and tested three classifiers on datasets collected from Saudi Twitter. Given that Saudi tweets represent two linguistic classes: MSA and SD, the task was to discriminate between Saudi dialect SD and MSA using supervised machine learning algorithms. We trained three classifiers: Logistic regression (LR), multi-nominal Naïve Bayes (MNB), and support vector machine (SVM) on a dataset that was automatically labeled as MSA or SD. The model for each configuration was built using two levels of language models (un-gram and bi-gram), as features for training. The systems reported high-accuracy performance (average 98.98%) when they were tested. However, when we tested these classifiers on another manually- annotated dataset and compared their results to automatic annotation, the best performance was reported by MNB achieving accuracy of 89.05. The drop in the performance probably occurred as a result of the factors mentioned previously in the discussion.

These results can be considered a baseline for future work. We look for improving our model by using additional feature sets and higher levels of language models. Moreover, we may use orthographic normalization tools that may have positive impacts on our model. As indicated, named entities represent another challenge for dialects identification. Hence, using NER tools can also be powerful for such a task. We seek to improve our systems knowing that SD identification is a significant step for learning lexical and morphological knowledge of this dialect, which can be beneficial for further Arabic NLP downstream tasks.

ACKNOWLEDGEMENT

The authors are grateful for the financial support received from King Khalid University for this research Under Grant No. R.G.P2/100/41.

REFERENCES

[1] Omar F Zaidan and Chris Callison-Burch. “Arabic dialect identification”. Computational Linguistics 40, pp. 171–202, 2014.
[2] Tommi Jauhiainen et al. “Automatic language identification in texts: A survey”. Journal of Artificial Intelligence Research 65, pp. 675–782, 2019.
[3] Charles A Ferguson. “Diglossia”. word 15, pp. 325–340, 1959.
[4] Heba Elfardy and Mona Diab. “Sentence level dialect identification in Arabic”. In: Proceed- ings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume2: Short Papers). 2013. Pp. 456–461.
[5] Abdulhadi Shoufan and Sumaya Alameri. “Natural language processing for dialectical Ara- bic: A Survey”. In: Proceedings of the second workshop on Arabic natural language process- ing. 2015. Pp. 36–48.
[6] Imane Guellil et al. “Arabic natural language processing: An overview”. Journal of King Saud University-Computer and Information Sciences, 2019.

[7] Kareem Darwish et al. “A panoramic survey of natural language processing in the Arab world”. Communications of the ACM 64, pp. 72–81, 2021.
[8] Christoph Tillmann, Saab Mansour, and Yaser Al-Onaizan. “Improved sentence-level arabic dialect classification”. In: Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects. 2014. Pp. 110–119.
[9] Mohamed Al-Badrashiny, Heba Elfardy, and Mona Diab. “Aida2: A hybrid approach for token and sentence level dialect identification in arabic”. In: Proceedings of the Nineteenth Conference on Computational Natural Language Learning. 2015. Pp. 42–51.
[10] Fatiha Sadat, Farnazeh Kazemi, and Atefeh Farzindar. “Automatic identification of arabic dialects in social media”. In: Proceedings of the first international workshop on Social media retrieval and analysis. 2014. Pp. 35–40
[11] Salima Harrat et al. “Cross-dialectal arabic processing”. In: International Conference on Intelligent Text Processing and Computational Linguistics. Springer. 2015. Pp. 620–632.
[12] Mohamed Lichouri et al. “Word-Level vs Sentence-Level Language Identification: Applica- tion to Algerian and Arabic Dialects”. Procedia Computer Science 142, pp. 246–253, 2018.
[13] Shervin Malmasi et al. “Discriminating between similar languages and arabic dialect identification: A report on the third dsl shared task”. In: Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3). 2016. Pp. 1–14.
[14] Shervin Malmasi, Eshrag Refaee, and Mark Dras. “Arabic dialect identification using a parallel multidialectal corpus”. In: Conference of the Pacific Association for Computational Linguistics. Springer. 2015. Pp. 35–53.
[15] Leena Lulu and Ashraf Elnagar. “Automatic Arabic dialect classification using deep learning models”. Procedia computer science 142, pp. 262–269, 2018.
[16] Mohamed Ali. “Character level convolutional neural network for Arabic dialect identifica- tion”. In: Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018). 2018. Pp. 122–127.
[17] Faisal Alshargi et al. “Morphologically Annotated Corpora for Seven Arabic Dialects: Taizi, Sanaani, Najdi, Jordanian, Syrian, Iraqi and Moroccan”. In: Proceedings of the Fourth Arabic Natural Language Processing Workshop. 2019. Pp. 137–147.
[18] Mohamed Elaraby and Muhammad Abdul-Mageed. “Deep models for arabic dialect identification on benchmarked data”. In: Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018). 2018. Pp. 263–274.
[19] Muhammad Abdul-Mageed, Hassan Alhuzali, and Mohamed Elaraby. “You tweet what you speak: A city-level dataset of arabic dialects”. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 2018.
[20] Mohammad Salameh, Houda Bouamor, and Nizar Habash. “Fine-grained arabic dialect identification”. In: Proceedings of the 27th International Conference on Computational Lin- guistics. 2018. Pp. 1332–1344.

AUTHORS

YAHYA ASERI is an assistant professor in Linguistics and Human language Technology in Arabic Language department at King khalid University. He serves as a consultant at Center for Artificial intelligence at KKU University, and he is also a member of ACL SIGARAB (ACL Special Interest Group on Arabic Natural Language Processing. He obtained his PhD degree in Linguistics from University of Colorado-Boulder, USA. His research interests focus on theoretical computational linguistics and its applications to hu- man language technology; namely, developing linguistic models for human language understanding, building annotated resources /corpora for machine learning, and developing natural language processing applications.

KHHALID ALREEMY is a software engineer with experiences in Artificial Intelligence, natural language processing and computer vision using deep learning models. He has a long experience working on C# for Microsoft and NET applications. He worked as a SQL database designer, specifically Oracle and MySQL exploiting the power of Python programming language for handling various data types. I also worked for diverse AI applications such as spam detection, sentiment analysis, object detection, Chatbots, and abnormality detection in mammogram images, etc. Currently he is working on various projects of Artificial Intelligence with highly oriented skills for data processing and preparation.

SALEM ALELYANI has been an assistant professor in the Computer Science Department at King Khalid University, Saudi Arabia, since 2014. He serves as a consultant to the president of the university and the director of the Center for Artificial Intelligence. He obtained his Ph.D. from Arizona State University in 2013 in Machine Learning and Data Mining. He has several publications in the field. He serves as a reviewer in multiple international conferences and scientific journals including ICTAI, AAAI, ICMLA, ICML, IRI, IEEE Access, Artificial Intelligence Review, AJSE, Information Sciences, and others. Also, he serves as a board member and as a PC member in other international journals and conferences.

MOHAMED MOHANA is a Mechatronics Engineer with a solid background in Control, Electronics, Mechanical, and Computer systems. He has done his Master's Engineering in Control and Automation using Artificial Intelligence; besides, He won the best master research from IEEE Malaysia Control System. He has four years of experience in Computer Vision and IoT, as well as MATLAB and Python programming languages. He has been involved in real-life industrial problems and overcame the challenges using IoT and Robots controlled by AI. He has created a UAV with an autopilot controller and a state-of-art IoT and Computer Vision device for potent security purposes. Recently, he is working with AI in renewable energy. However, he has the ability to see the whole picture from solving the problem (Research) up to the solution deployment (Development), to create Artificial Intelligence solutions and products for real-life situations.