Classifying Arabic dialect text in the Social Media Arabic Dialect Corpus (SMADC)

Areej Alshutayri
College of Computer Science and Engineering
University of Jeddah
Jeddah, Saudi Arabia
aoalshutayri@uj.edu.sa

Eric Atwell
School of Computing
University of Leeds
Leeds, United Kingdom
e.s.atwell@leeds.ac.uk

Abstract

In recent years, research in Natural Language Processing (NLP) on Arabic has garnered significant attention. This includes research about classification of Arabic dialect texts, but due to the lack of Arabic dialect text corpora this research has not achieved a high accuracy. Arabic dialects text classification is becoming important due to the increasing use of Arabic dialect in social media, so this text is now considered quite appropriate as a medium of communication and as a source of a corpus. We collected tweets, comments from Facebook and online newspapers representing five groups of Arabic dialects: Gulf, Iraqi, Egyptian, Levantine, and North African. This paper investigates how to classify Arabic dialects in text by extracting lexicons for each dialect which show the distinctive vocabulary differences between dialects. We describe the lexicon-based methods used to classify Arabic dialect texts and present the results, in addition to techniques used to improve accuracy.

1 Introduction

Textual Language Identification or Dialect Identification is the task of identifying the language or dialect of a written text. The Arabic language is one of the world's major languages, and it is considered the fifth most-spoken language and one of the oldest languages in the world. Additionally, the Arabic language consists of multiple variants, both formal and informal (Habash, 2010). Modern Standard Arabic (MSA) is a common standard written form used worldwide. MSA is derived from Classical Arabic which is based on the text of the Quran, the holy book of Islam; MSA is the primary form of the Arabic language that is spoken and studied today. MSA is taught in Arab schools, and promoted by Arab civil as well as religious authorities and governments. There are many dialects spoken around the Arab World; Arabic dialectologists have studied hundreds of local variations, but generally agree these cluster into five main regional dialects: Iraqi Dialect (IRQ), Levantine Dialect (LEV), Egyptian Dialect (EGY), North African Dialect (NOR), and Gulf Dialect (GLF). Arabic dialectologists have traditionally focused mainly on variation in phonetics or pronunciation of spoken Arabic; but Arabic dialect text classification is becoming important due to the increasing use of Arabic dialect in social media text. As a result, there is a need to know the dialect used by Arabic writers to communicate with each other; and to identify the dialect before machine translation takes place, in order to ensure spell checkers work, or to accurately search and retrieve data. Furthermore, identifying the dialect may improve the Part-Of-Speech tagging: for example, the MADAMIRA toolkit identifies the dialect (MSA or EGY) prior to the POS tagging (Pasha et al., 2014). The task of Sentiment Analysis of texts, classifying the text as positive or negative sentiment, is also dialect-specific, as some diagnostic words (especially negation) differ from one dialect to another. Text classification is identifying a predefined class or category for a written document by exploring its characteristics or features (Ikonomakis et al., 2005; Sababa and Stassopoulou, 2018). However, Arabic dialect text classification still needs a lot of research to increase the accuracy of classification due to the same characters being used to write MSA text and dialects, and also because there is no standard written format for Arabic dialects. This paper sought to find appropriate lexical fea-

tures to classify Arabic dialects and build a more sophisticated filter to extract features from Arabic-character written dialect text files. In this paper, the corpus was annotated with dialect labels and used in automatic dialect lexicon-extraction and text-classification experiments.

2 Related Work

There are many studies that aim to classify Arabic dialects in both text and speech; most spoken Arabic dialect research focuses on phonological variation and acoustic features, based on audio recordings and listening to dialect speakers. In this research, the classification of Arabic dialects will focus on text, One example project focused on Algerian dialect identification using unsupervised learning based on a lexicon (Guellil and Azouaou, 2016). To classify Algerian dialect the authors used three types of identification: total, partial and improved Levenshtein distance. The total identification meant the term was present in the lexicon. The partial identification meant the term was partially present in the lexicon. The improved Levenshtein applied when the term was present in the lexicon but with different written form. They applied their method on 100 comments collected from the Facebook page of Djezzy and achieved an accuracy of 60%. A lexicon-based method was used in (Adouane and Dobnik, 2017) to identify the language of each word in Algerian Arabic text written in social media. The research classified words into six languages: Algerian Arabic (ALG), Modern Standard Arabic (MSA), French (FRC), Berber (BER), English (ENG) and Borrowings (BOR). The lexicon list contains only one occurrence for each word and all ambiguous words which can appear in more than one language are deleted from the list. The model was evaluated using 578 documents and the overall accuracy achieved using the lexicon method is 82%. Another approach to classify Arabic dialect is using text mining techniques (Al-Walaie and Khan, 2017). The text used in the classification was collected from Twitter. The authors used 2000 tweets and the classification was done on six Arabic dialects: Egyptian, Gulf, Shami, Iraqi, Moroccan and Sudanese. To classify text, decision tree, Naïve Bayes, and rule-based Ripper classification algorithms were used to train the model with keywords as features for distinguishing one dialect from another, and to test the model the used 10-fold cross-validation. The best accuracy scored 71.18% using rule-based (Ripper) classifier, 71.09% using Naïve Bayes, and 57.43% using decision tree. Other researchers on Arabic dialect classification have used corpora limited to a subset of dialects; our SMADC corpus is an International corpus of Arabic with a balanced coverage of all five major Arabic dialect classes.

3 Data

The dataset used in this paper is the Social Media Arabic Dialect Corpus (SMADC) which was collected using Twitter, Facebook and comments from online newspapers described in (Alshutayri and Atwell, 2017, 2018b,c). We plan to make the Social Media Arabic Dialect Corpus (SMADC) available to other researchers for non commercial uses, in two formats (raw and cleaned) and with a range of metadata. This corpus covers all five major Arabic dialects recognised in the Arabic dialectology literature: EGY, GLF, LEV, IRQ, and NOR. Therefore, five dictionaries were created to cover EGY dialect, GLF dialect, LEV dialect, IRQ dialect, and NOR dialect. (Alshutayri and Atwell, 2018a) presented the annotation system or tool which was used to label every document with the correct dialect tag. The data used in the lexicon based method was the result of the annotation, and each comment/tweet is labelled either dialectal document or MSA document.

The MSA documents in our labelled corpus were used to create an MSA word list, then we added to this list MSA stop words collected from Arabic web pages by Zerrouki and Amara (2009), and the MSA word list collected from Sketch Engine (Kilgarriff et al., 2014), in addition to the list of MSA seed words for MSA web-as-corpus harvesting, produced by translating an English list of seed words (Sharoff, 2006). The final MSA word list contains 29674 words. This word list is called "StopWords1" and was used in deleting all MSA words from dialect documents, as these may contain some MSA words, for example due to code switching between MSA and dialect.

The dialectal documents consist of documents and dialectal terms, where the annotators (players) were asked to write the dialectal terms in each document which help them to identify dialect as described in (Alshutayri and Atwell, 2018a). The dialectal documents were divided into two sets: 80% of the documents were used to create dialectal dic-

tionaries for each dialect, and 20%, the rest of the documents, were used to test the system. To evaluate the performance of the lexicon based models, a subset of 1633 documents was randomly selected from the annotated dataset and divided into two sets; the training dataset which contains 1383 documents (18,697 tokens) are used to create the dictionaries, and the evaluation dataset which contains 250 documents (7,341 tokens). The evaluation dataset did not include any document used to create the lexicons as described previously.

4 Lexicon Based Methods

To classify the Arabic dialect text using the Lexicons, we used a range of different classification metrics and conducted five experiments, all of which used a dictionary for each dialect. The following sections show the different methods used and describe the difference between the conducted experiments, and the result of each experiment.

4.1 Dialectal Terms Method

In this method, the classification process starts at the word level to identify and label the dialect of each word, then the word-labels are combined to identify the dialect of the document. The dialectal terms produced from the annotation tool were used as a dictionary for each dialect. The proposed system consists of five dictionaries, one for each dialect: EGY dictionary contains 451 words, GLF dictionary contains 392 words, IRQ dictionary contains 370 words, LEV dictionary contains 312 words from LEV, and NOR dictionary contains 352 words.

According to the architecture in Figure 1, to classify each document as being a specific dialect, the system follows four steps:

- Detect the MSA words in the document by comparing each word with the MSA words list, then delete all MSA words found in the document.
- The result from the first step is a document containing only dialectal words.
- Detect the dialect for each word in the document by comparing each word with the words in the dictionaries created for each dialect.
- Identify dialect.

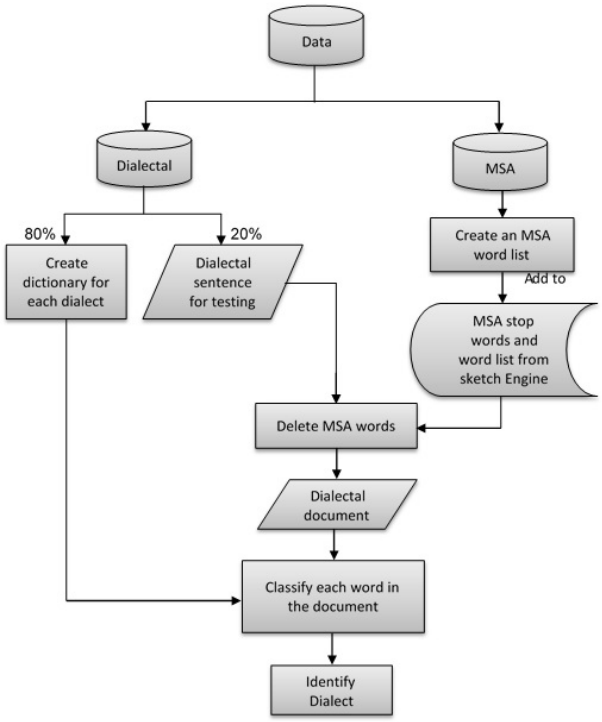

Figure 1: The architecture of classification process using lexicon based.

Using this method based on the dialectal terms written by the annotators produces some unclassified documents due to words that occur in more than one dialect. For example, the document in Figure 2 was labelled as LEV and the structure of the document is also LEV dialect, but the word كتير (kti:r) which appears in the text is also used in EGY. Therefore, when classifying each word in the document the model found the word كتير (kti:r) in EGY dictionary and also in LEV dictionary, so the model was not able to classify this document as the other words are MSA words or shared dialectal words. Unclassified documents indicate that using this dialectal terms method is not effective in dealing with ambiguous words.

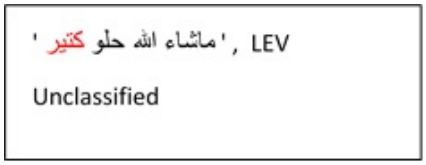

Figure 2: Example of unclassified document.

Table 1 shows the accuracies achieved by applying the dialectal terms method on the testing set. The first column represents using MSA words list, and the second column represents the achieved ac-

curacies based on using SMADC to create dictionaries. The best accuracy is 56.91 with 140 documents correctly classified using StopWords1. Based on this method, 110 documents were unclassified to a specific dialect because they contain some ambiguous terms which are used in more than one dialect, as in the example of Figure 2. As a solution to this problem, a voting method is used and another way is using a frequent term method.

MSA	SMADC
StopWords1	**56.91**%
Without delete MSA Words	55.60%

Table 1: Results of dialectal terms method using the dictionaries created from SMADC.

4.2 Voting Methods

Another method to classify Arabic dialect text is to treat the text classification of Arabic dialects as a logical constraint satisfaction problem. The voting method is an extension of the dialectal term method presented previously. The classification starts at the word level based on the dictionaries created from the 80% training set of documents described in Section 3. So, the annotated training set of documents was used instead of the dialectal terms list. In this method, we looked to the whole document and count how many words belong to each dialect. Each document in the voting method was represented by a matrix C. The size of the matrix is $C_{|n|\times|5|}$, where n is the number of words in each document, and 5 is the number of dialects (EGY, NOR, GLF, LEV, and IRQ).

4.2.1 Simple Voting Method

In this method, the document is split into words and the existence of each word in the dictionary is represented by 1 as in Equation 1.

$$c_{ij} = \begin{cases} 1, & \text{if } \ word \ \ \epsilon \ dialect \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

The following illustrates the method. We apply Equation 1 on the following document **A** labelled as IRQ dialect as in Figure 3:

The result of classification is IRQ according to Table 2; the total shows that four words in this document belong to IRQ dialect in comparison with two words belong to NOR and EGY, and one word belong to LEV and GLF.

يعجبني اغرد عن كلشي يخطر بالي ,IRQ

Translation: I like to tweet about anything come to my mind

Figure 3: The text in document **A**.

Words	NOR	EGY	IRQ	LEV	GLF
يعجبني	0	0	1	0	0
اغرد	0	0	1	0	0
عن	1	1	1	1	1
كلشي	1	0	1	0	0
يخطر	0	0	0	0	0
بَالي	0	1	0	0	0
Total	2	2	**4**	1	1

Table 2: The matrix representation of document **A** with simple voting.

The proposed model identifies the document correctly but sometimes this model cannot classify a document and the result is unclassified when more than one dialect gets the same count of words (total), like document **B** labelled as GLF dialect as shown in Figure 4:

هههههههههه خليتني اضحك من قلب ليش تتكلم على زوجتك بهالطريقة لاحظتك معلق على موضوعين بس أقول الله يعينك للحين في حريم تتصرف بهالشكل ,GLF

Translation: Hhhhhhhhhhh you made me laughing hard why you talking about your wife in this way, I noticed you commenting on two topics but I say God helps you, until now there are women behave like this.

Figure 4: The text in document **B**

Using the StopWords1 to delete MSA words from the document, the result is the following dialectal document containing only dialectal words as in Figure 5.

According to the result in Table 3 the document is unclassified because more than one dialect has the same total number of words.

4.2.2 Weighted Voting Method

This method is used to solve the problem of unclassified documents in simple voting method. To solve this problem, we proposed to change the value of the word from 1 to the probability of the word to belong to this dialect as a fraction of one divided by the number of dialects the word is found in their dictionaries as in Equation 2. If a word can belong to more than one dialect, its vote

هههههههههه خليتني ليش بهالطريقة بس للحين بهالشكل

Figure 5: Example of unclassified document.

Words	NOR	EGY	IRQ	LEV	GLF
هّههههّه	0	1	1	1	0
خليتني	0	0	0	0	0
ليش	1	0	1	1	1
بهَالطريقة	0	0	0	0	0
بس	1	1	1	1	1
للّحين	0	0	0	0	1
بهَالشكل	0	0	0	0	0
Total	2	2	3	3	3

Table 3: The matrix representation of document **B** with simple voting.

is shared between the dialects.

$$c_{ij} = \begin{cases} \frac{1}{m}, & \text{if } word \ \epsilon \ dialect \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

$\frac{1}{m}$ is the probability of the word belonging to the specific dialect, where m the number of dialects which the word belongs to. By applying the new method on the unclassified document, the document is classified correctly as GLF dialect, according to Table 4.

Words	NOR	EGY	IRQ	LEV	GLF
هّههههّه	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0
خليتني	0	0	0	0	0
ليش	$\frac{1}{4}$	0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
بهَالطريقة	0	0	0	0	0
بس	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$
للّحين	0	0	0	0	1
بهَالشكل	0	0	0	0	0
Total	0.45	0.5333	0.7833	0.7833	**1.45**

Table 4: The matrix representation of document **B** with Weighted voting.

4.2.3 Results of Voting Method

Voting method is focused on the existence of the word in the dictionary, so, the frequency of the word is ignored, unlike the frequent term method which described in Section 4.3. The highest accuracy achieved is 74.0% without deleting MSA words from the classified document. Moreover, using the value of one to express the existence of the word in the dictionary showed low accuracy due to the similarity between the sum of ones for each dialect, as described in Section 4.2.1. Table 5 shows the different accuracies achieved using SMADC. The first column in Table 5 shows using of MSA stop words. The second and the third columns represent the methods used to classify documents. The cells inside the second and third columns present the achieved accuracies using these methods. The voting method scored 74% using the weighted voting method and SMADC to create dictionaries. After cleaning the MSA word list, the accuracy increased to 77.60%.

MSA	Simple Vote	Weighted Vote
StopWords1	69.19%	**72.0%**
Without delete MSA Words	65.60%	**74.0%**

Table 5: Results of Voting methods using the dictionaries created from SMADC.

4.3 Frequent Terms Methods

Another method is presented in this section to solve the problem shown in the dialectal terms method described in Section 4.1 and to improve the accuracy of classification achieved using the voting method. In frequent terms method, new dictionaries with word frequencies were created from the 80% training set of documents. The documents were classified into the five dialects. Then, for each dialect a .txt file was created to contain one word per line with the word's frequency based on the number of times the word appeared in the documents. The frequency for each word showed if the word is frequent in this dialect or not, which helps to improve the accuracy of the classification process. In comparison to the first method, the third step in Figure 1 was used to detect the dialect for each word in the document by comparing each word with the words in the dictionaries created for each dialect. If the word is in the dictionary, then calculate the weight (W) for each word by dividing the word's frequency (F) value by the Length of the dictionary (L) which equals the total number of words in the word's dialect dictionary, using the following equation:

$$W(word, dict) = \frac{F(word)}{L(dict)} \quad (3)$$

For each document, five vectors were created, one per dialect, to store the weight for each word in the document; so the length of each vector is equal to the length of the document. By applying the Equation 3 on "كتير", we found the weight of the word "كتير" in LEV dialect is bigger than the weight of it in EGY dialect, as shown in the following equations.

$$W(\text{"كتير"}, EGY) = \frac{F(\text{"كتير"})}{L(EGY)} = \frac{3}{2032} = 0.00147$$

$$W(\text{"كتير"}, LEV) = \frac{F(\text{"كتير"})}{L(LEV)} = \frac{8}{2028} = 0.00394$$

Two experiments were done after calculating the weight for each word. The first experiment was based on summing the weights and calculating the average. The second experiment was based on multiplying the weights together.

4.3.1 Weight Average Method (WAM)

This method based on calculating the average of the word weights for each document. Table 6 shows the values of the weight for each word in the document after deleting MSA words. Five vectors were created to represent five dialects and each cell contains the weight for each word in the document. The model calculated the average for each dialect by taking the summation of the weight (W) values for each vector then dividing the summation of weights by the length (L) of the document after deleting the MSA words, as in the following equation:

$$Avg_{dialect} = \frac{\sum W_{dialect}}{L(document)} \quad (4)$$

Words	NOR	LEV	IRQ	GLF	EGY
مَاشَاء	0	0.00049309	0	0.00026143	0
حلو	0	0.00295857	0.00053304	0.00026143	0.00049212
كتير	0	0.00394477	0	0	0.00147637

Table 6: Results of WAM using the dictionaries created from SMADC.

By calculating the average for the dialect vectors using the Equation 4, the model classified the document as LEV dialect, after comparing the results of the average obtained from the following equations.

$$Avg_{EGY} = \frac{\sum W_{EGY}}{L(document)} = \frac{0.00196849}{3} = 0.00065616$$

$$Avg_{LEV} = \frac{\sum W_{LEV}}{L(document)} = \frac{0.00739643}{3} = 0.00246547$$

$$Avg_{GLF} = \frac{\sum W_{GLF}}{L(document)} = \frac{0.00052286}{3} = 0.00017428$$

$$Avg_{IRQ} = \frac{\sum W_{IRQ}}{L(document)} = \frac{0.00053304}{3} = 0.00017768$$

By applying the proposed model on the same unclassified example in Figure 2, we found that the model classified the document correctly as in Figure 6.

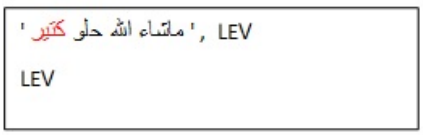

Figure 6: Example of correctly classified document.

4.3.2 Weight Multiplied Method (WMM)

The WAM model is based on summing the word weights and calculating the average. According to probability theory, probabilities are generally combined by multiplication. So, for an alternative model, the Weight Multiplied Method (WMM), we multiplied the word weights for each document to compute the accuracy of classification in comparison to the average method used in the previous section.

$$P(doc|c) = \prod W(word, dict) \quad (5)$$

We applied Equation 5 on the weights in Table 6. There is a problem with combining weights by multiplication: if any of the weights to be combined is zero, the combined weight will be zero. So, we change the value of not found words in the dialect dictionary from zero to one. However, in the Table 6 if the values in NOR vector changed to one this will affect the result of multiplication. For that reason the result of multiplication was checked as to whether or not it equal one then we changed the result to zero.

According to Equation 5 the document is classified as IRQ dialect, which is a wrong prediction.

$P_{EGY} = \prod W(word|EGY) = 1 \times 0.00049212 \times 0.00147637 = 0.00000072$

$P_{LEV} = \prod W(word|LEV) = 0.00049309 \times 0.00295857 \times 0.00394477 = 0.0000000057$

$P_{GLF} = \prod W(word|GLF) = 0.00026143 \times 0.00026143 \times 1 = 0.000000068$

$P_{IRQ} = \prod W(word|IRQ) = 1 \times 0.00053304 \times 1 = 0.00053304$

To solve wrong predictions which result from using WMM and to improve the classification accuracy, we replace one when the word is not in the dictionary with one divided by the number of words in each dictionary to not affect the result of multiplication. By applying the new value to Equation 5 the document is correctly classified as LEV dialect. Table 7 shows the improved accuracy resulted using WMM when using one divided by the number of words in each dictionary to represent the absence of a word in the dictionary.

$P_{EGY} = \prod W(word|EGY) = \frac{1}{L(dic_{EGY})} \times 0.00049212 \times 0.00147637 = \frac{1}{2032} \times 0.00049212 \times 0.00147637 = 0.0000000003575$

$P_{LEV} = \prod W(word|LEV) = 0.00049309 \times 0.00295857 \times 0.00394477 = 0.0000000057$

$P_{GLF} = \prod W(word|GLF) = 0.00026143 \times 0.00026143 \times \frac{1}{L(dic_{GLF})} = 0.00026143 \times 0.00026143 \times \frac{1}{3472} = 0.000000068$

$P_{IRQ} = \prod W(word|IRQ) = \frac{1}{L(dic_{IRQ})} \times 0.00053304 \times \frac{1}{L(dic_{IRQ})} = \frac{1}{1889} \times 0.00053304 \times \frac{1}{1889} = 0.0000000000149$

$P_{NOR} = \prod W(word|NOR) = \frac{1}{L(dic_{NOR})} \times \frac{1}{L(dic_{NOR})} \times \frac{1}{L(dic_{NOR})} = \frac{1}{1436} \times \frac{1}{1436} \times \frac{1}{1436} = 0.0000000003376$

The following sections will compare the first model based on summation and calculate average with the multiplication method, and show the achieved results using average method and the multiplication method.

4.3.3 Result of Frequent Terms Method

The frequent term method which is based on using word frequencies gave good results in showing whether the words in the tested document is used in the specific dialect. The model was tested using the test dataset described in Section 3. Based on the average method, the model achieved 88% accuracy using the MSA StopWords1 list. However, using the multiply method achieves low accuracy due to replacing zero with one when the word does not exist in the dictionary; so instead we divided by the number of words in the dictionary..

Table 7 reports the different accuracies achieved when using SMADC based on using one divided by the number of words in the dictionary to represent words which are not found in the dictionary. The frequent terms method scored 88% using the weight average metric when dictionaries were created using SMADC. The accuracy improved to 90% after cleaning the MSA word list from some dialectal words as a result of mislabelling process.

MSA	WMM	WAM
StopWords1	55.60%	**88.0%**
Without delete MSA Words	43.0%	**64.0%**

Table 7: Results of frequent term methods using the dictionaries created from SMADC.

By comparing the Weight Average Method (WAM) model based on summation and calculating average with the Weight Multiplied Method (WMM), we found that the WAM achieved a higher accuracy than the WMM multiplication method.

5 Conclusion

The classification of Arabic dialect text is a hot topic attracting a number of studies over the last ten years (Sadat et al., 2014; Zaidan and Callison-Burch, 2014; Elfardy and Diab, 2013; Mubarak and Darwish, 2014; Harrat et al., 2014; Shoufan and Alameri, 2015). In this paper, we classified Arabic dialects text using a lexicon based method, and explored different metrics for scoring dialect words from lexicons: weight average method, weight multiplied method, simple voting method and weighted voting method. The lexicons were dictionaries created for each dialect

from our Arabic dialect corpora. The classification process was based on deleting all MSA words from the document then checking each word in the document by searching the dialect dictionaries. The voting method scored 74% using the weighted voting method and SMADC to create dictionaries. After cleaning the MSA word list, the accuracy increased to 77.60%. The frequent terms method scored 88% using the weight average metric when dictionaries were created using SMADC. The accuracy improved to 90% after cleaning the MSA word list from some dialectal words as a result of mislabelling process. These scores compare favourably against other Arabic dialect classification research on subsets of Arabic dialects.

References

Wafia Adouane and Simon Dobnik. 2017. Identification of languages in Algerian Arabic multilingual documents. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 1–8, Valencia, Spain. Association for Computational Linguistics.

Mona Al-Walaie and Muhammad Khan. 2017. Arabic dialects classification using text mining techniques. In *2017 International Conference on Computer and Applications (ICCA)*, pages 325–329.

Areej Alshutayri and Eric Atwell. 2017. Exploring twitter as a source of an arabic dialect corpus. *International Journal of Computational Linguistics (IJCL)*, 8:37–44.

Areej Alshutayri and Eric Atwell. 2018a. Arabic dialects annotation using an online game. In *2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–5.

Areej Alshutayri and Eric Atwell. 2018b. Creating an arabic dialect text corpus by exploring twitter, facebook, and online newspapers. In *Proceedings of Open-Source Arabic Corpora and Processing Tools. OSACT'2018 Open-Source*, Miyazaki, Japan.

Areej Alshutayri and Eric Atwell. 2018c. *A Social Media Corpus of Arabic Dialect Text*. Computer-Mediated Communication and Social Media Corpora, Clermont-Ferrand: Presses universitaires Blaise Pascal.

Heba Elfardy and Mona Diab. 2013. Sentence level dialect identification in Arabic. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 456–461, Sofia, Bulgaria. Association for Computational Linguistics.

Imène Guellil and Faiçal Azouaou. 2016. Arabic dialect identification with an unsupervised learning (based on a lexicon). application case: Algerian dialect. In *IEEE Intl Conference on Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES)*, pages 724–731.

Nizar Y. Habash. 2010. *Introduction to Arabic Natural Language Processing*. Morgan and Claypool.

Salima Harrat, Karima Meftouh, Mourad Abbas, and Kamel Smaïli. 2014. Building resources for algerian arabic dialects. In *15th Annual Conference of the International Communication Association (Interspeech)*, pages 2123–2127.

Emmanouil Ikonomakis, Sotiris Kotsiantis, and V Tampakas. 2005. Text classification using machine learning techniques. *WSEAS transactions on computers*, 4:966–974.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, 1(1):7–36.

Hamdy Mubarak and Kareem Darwish. 2014. Using twitter to collect a multi-dialectal corpus of Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 1–7, Doha, Qatar. Association for Computational Linguistics.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1094–1101, Reykjavik, Iceland. European Language Resources Association (ELRA).

Hanna Sababa and Athena Stassopoulou. 2018. A classifier to distinguish between cypriot greek and standard modern greek. pages 251–255.

Fatiha Sadat, Farnzeh Kazemi, and Atefeh Farzindar. 2014. Automatic identification of Arabic language varieties and dialects in social media. In *Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP)*, pages 22–27, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Serge Sharoff. 2006. Open-source corpora: Using the net to fish for linguistic data. *International Journal of Corpus Linguistics*, 11:435–462.

Abdulhadi Shoufan and Sumaya Alameri. 2015. Natural language processing for dialectical arabic: A survey. In *Proceedings of the Second Workshop on*

Arabic Natural Language Processing, pages 36–48, Beijing, China. Association for Computational Linguistics.

Omar F. Zaidan and Chris Callison-Burch. 2014. Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.