Arabic dialects identification: North African dialects case study

Mohamed Berrimi[a], Abdelouahab Moussaoui[a], Mourad Oussalah[b] and Mohamed Saidi[a]

[a]Department of computer sciences, University of Ferhat Abbas 1, Algeria
[b]Department of Computer Science and Engineering, University of Oulu, Finland

Abstract

Arabic is the fourth most used language on the Internet and the official language of more than 20 countries around the world. It has three main varieties, Modern Standard Arabic, which is used in books, news and education, local Dialects that vary from region to another, and Classical Arabic, the written language of the Quran. Maghrebi dialect is the Arabic dialect language used in North African countries, where internet users from these countries feel more comfortable using local slangs than native Arabic. In this study, we present a large dataset of regional dialects of three countries, namely Algeria, Tunisia, and Morocco, then we investigate the identification of each dialect using a machine learning classifiers with TF-IDF features. The approach shows promising results, where we achieved accuracy up to 96%.

Keywords

Arabic dialects, Arabic text processing, Feature extraction, Text classification

1. Introduction

Arabic is the fourth most used language on the Internet with more than 400 million Arabic speakers [1], and the official language of 22 countries[2]. It presents severe challenges to researchers due to its particular format. Arabic is a highly structured and derivational language where morphology plays a significant role, and has three main varieties from Modern Standard Arabic MSA, Arabic Dialect or local slangs, and Classical Arabic [1].

MSA is the formal Arabic that is used in news broadcasting channels, books, and in education, Which is characterized by grammar rules. Unlike MSA, a regional dialect does not have a specific written set of grammar rules regulated by an authoritative organization, and it's only used in informal communication.

Sabbagh et al. [3] presented a list of different Arabic dialects along with their regions, where Egyptian dialect is considered as the most widely understood within Arabic countries due to the Egyptian cinema industry, which is very popular in the Arab world.

IAM'20: Third conference on informatics and applied mathematics, 21–22 October 2020,Guelma, ALGERIA
mohamed.berrimi@univ-setif.dz (M. Berrimi); Abdelouahab.Moussaoui@univ-setif.dz (A. Moussaoui); mourad.oussalah@oulu.fi (M. Oussalah); Mohamed.Saidi@univ-setif.dz (M. Saidi)
0000-0003-4678-5328 (M. Berrimi)

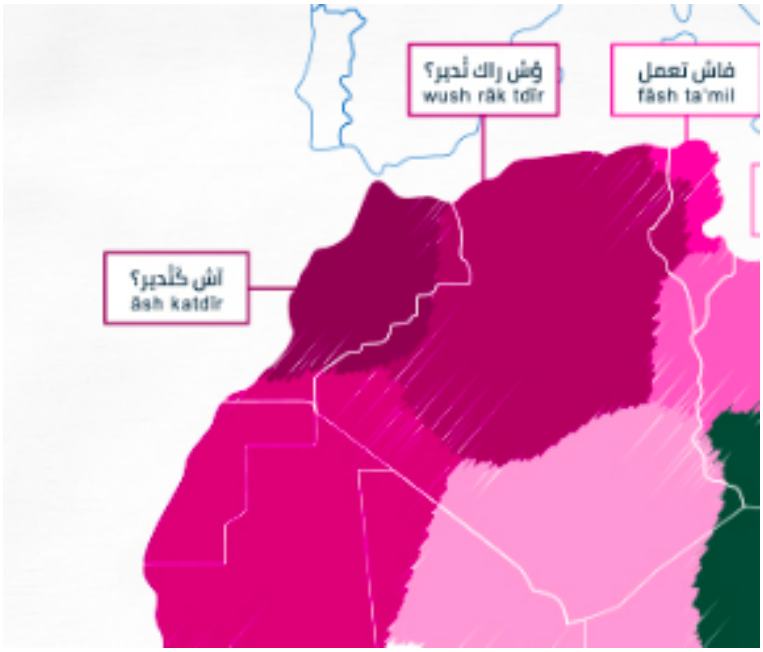

Figure 1: Different dialects in the Maghreb countries, where people living in the borders speak a similar dialect to the country's neighbours, the graphic shows how to say '*what are you doing*' in different dialects[8].

Levantine covers a group of spoken dialects along with Palestine, Syria, Lebanon, and Jordan, with more than 30 million speakers worldwide[4].

Gulf slang is the closed regional dialect to MSA, Iraki, on the other side, is used in Irak and nearby regions. The French language heavily influenced Maghrebi due to French colonialism in the last century. Unlike Egyptian and levantine slangs, Maghrebi dialects are only understood within this region; this could be because these dialects contain French words.

In this study, we focus on the Maghrebi dialect, which appears more on social media, where users from these countries are usually less comfortable communicating in MSA than in their dialect. Also, they use Arabizi (majorly French-Arabizi), which is a form of Arabic text written with Latin characters.
Unlike the rest of Arab countries, Facebook is the most widely used website in Algeria 57.05% [5], Tunisia with 83% [6] and Morocco 43% [7].

Each of these countries has its dialect, but in Algeria, for instance, which is in the middle of the Maghreb world, Whenever you go to the east, you find people speaking Tunisian dialects. The same thing for the west side of the country where people speak Moroccan dialect. Furthermore, differentiating between these dialects becomes challenging.

These dialects are considered as low resource languages, where there's a lack of available data compared to native languages. French-Arabizi poses challenges on data scientists due to its unstructured syntax and doesn't follow any grammatical rules. It uses a mix of French and

Arabic words in a single sentence, like: 'Saha Mon frére, aprés nchallah nroho 3ndo' i.e., " Ok Brother, later, God willing, we will go to him." and varies among different regions [9].

In this paper, we present a large dialect dataset of the three north African countries, namely Algeria, Tunisia, and Morocco, and explore feature extraction techniques such as TF-IDF weighting, and train machine learning classifiers for the identification of each dialect.

The rest of the paper is organized as follows. In section 2, we discuss the recent work interested in the identification of Arabizi and Arabic dialects, also some proposed corpora. In section 3, we present the collected dataset with its annotation. In section 3, we demonstrate the different approaches and methodology followed to propose a good baseline for the detection of different dialects. Section 4 summarized the major findings of the research, and then we conclude in section 5, the use of machine learning-based algorithms on the Low resource languages and Arabizi.

2. Related works

recently more research focused on the identification of different Arabic dialects on social media, as well as the collection of data. Sayadi et al. [10] provided a manually annotated dataset with almost 50,000 tweets from 8293 users, then studied sentiment analysis on Tunisian dialect and Modern Standard Arabic.

Tobaili [11] annotated a corpus of the splitTwitter data stream coming from within Lebanon and Egypt, where users speak Araby-Englizi, then trained a classifier and achieved an average classification accuracy of 93% and 96% for Lebanon and Egypt datasets respectively.
Guellil et al. [12] proposed an approach for Arabic dialect identification in social media, specifically the Algeria dialect. The authors applied their approach to 100 messages manually annotated, and they achieved accuracy more than 60%.
Seddah et al. [13] introduced the first treebank for a romanized user-generated content variety of Algerian dialect, as mentioned in their paper. The content written in the Arabic language on the Internet is characterized by a high degree of linguistic diversity due to the use of colloquial dialects and writing in Roman characters, in addition to the phenomenon of code-switching. In addition to the annotated data, the authors provide around 1 million tokens (over 46k sentences) of unlabeled Arabizi content.

Darwish [14] addressed the problem of identifying Arabizi (Arabic text written with Latin characters) using word and sequence-level features achieving 98.5%, then convert it into Arabic characters using transliteration mining with language modeling achieving 88.7%

Many studies also focused on the collection of Arabizi and different Arabic dialects corpora from social media, Zaidan et al.[15] collected a corpus, from three Arabic newspapers of Levantine, Gulf, and Egyptian dialects.
Cotterell et al.[16] also presented extensive dialectal data from online resources for Algerian,

	Algerian	Moroccan	Tunisian
Number of text sequence	21230	20150	19050

Table 1
Size of data per class

Egyptian, Iraki, and Gulf.

In this work, we focus on the collection of North African (Maghreb) Dialects for Algerian, Moroccan, and Tunisian, and also training machine learning classifiers for the identification of different dialects.

3. Dataset

Facebook is the most popular social media website in North Africa. in North African countries; for this purpose, we searched for most popular Facebook pages (where the number of followers is higher than 100k) for Algerian, Tunisian, and Moroccan communities, where posts and comments are usually written in local dialects.

After grouping the Facebook pages per country, we manually collected around 20000 posts and comments on these pages, excluding the name of the commenters (only comment and post text body were scrapped). We labeled each group, resulting in a dataset with a total of 60000 text sequences with three main balanced classes.

4. Proposition

In this section, we detail the preprocessing phase, and different feature extraction methods as well as the experiments we carried on the cleaned dataset.

4.1. Preprocessing

Preprocessing is a technique that is used to convert the raw data into a clean one. Data collected from social media may contain special characters, words with repeated characters (like: Sahaaa Khoyaaa) , URLs, emoticons, punctuations, and unnecessary words. Accordingly, we applied cleaning functions to enhance the morphology of the presented text sequences and reduce the noise.

We eliminated numbers, URLs and the hashtags by deleting the # symbol, We also removed special characters like punctuations, emojis, Arabic diacritics, and words with two characters. After collection, we noticed that posts are in some cases too long, with more than ten lines. Thus, we splited sentences containing six words.

The results of different steps of preprocessing performed on our dataset is illustrated on table

	Algerian	Moroccan	Tunisian
Number of text sequence	21700	20550	19200

Table 2
Size of the dataset after splitting the long sentences into smaller ones

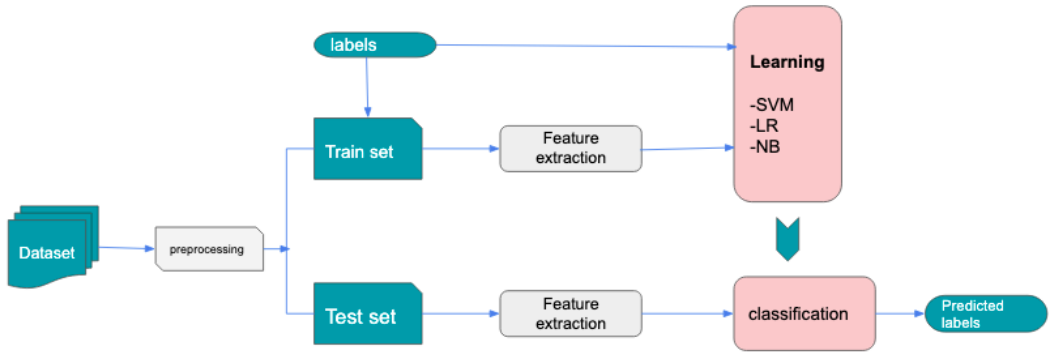

Figure 2: Generic graph showing the overall methodology

Algerian	وش يا خويا ak tconnecti b pc !!	Hey dude, are you connecting on the laptop!!
Morrocan	aslan la science est basée sur le partage, wakha a sahbi	Science is basically based on sharing, dude.
Tunisian	Barcha tmas5ir hnaya, allah yehlekom rana faykin bikom wino il pétrole.	Much nonsense, may Allah punish you all, we are aware of you, where's the oil!

Figure 3: Sample of sentences written in Arabizi.

Stop word removal: stop words are lists of words that occur much on textual data with no added meaning to the sentence.

Since Arabizi contains words in Arabic, French and English we removed all stop words lists from these languages, we also removed words that occur in the three classes, like persons and months names, personal pronouns like (salem, haya, sahbi, houwa, hiya, houma..etc) which are often repeated on the dataset, by performing this, we removed 798 terms.

Sample of removed words that belongs that are repeated in three classes After cleaning we made a split of 80% for the training set, and the rest was left for validating and testing the models.

	Text sequence	Translation
Algerian	معلابالیش کیفاش یخمو هذوك لعباد یا شکوبي, تقول معندهمش راس	I don't know how these people think, they are headless!
Morrocan	إندخلو غیر بشوي بایاك مزال ناعس و مریض منبغیوش نصدعوه صافي	Let's go inside quietly; your dad is sick we don't need to stress him more
Tunisian	المهم قالوا اكهو باش نروحوا للبلاد	The important thing is that they mentioned it, so we go home.

Figure 4: Sample of sentences written in local dialects.

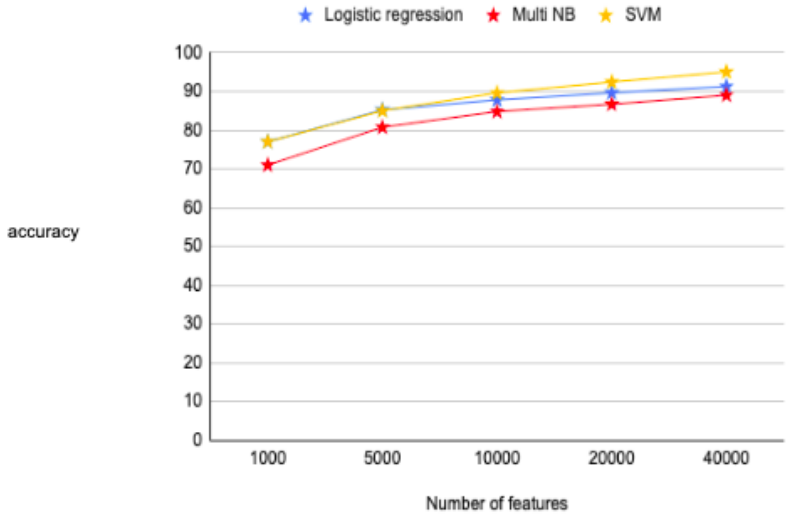

Figure 5: Accuracy level according to the number of features used in TF-IDF.

4.2. Feature extraction

In this study we have used TF-IDF, short for term frequency-inverse document frequency, commonly used to determine the importance of a specific term in the document.

The idea behind TF-IDF is to represent each word in a document by a number, or weight, that is proportional to its frequency (occurrences) in the document, and inversely proportional to the number of documents in which it occurs, meaning words that occurs the most within a document will end up having small weights, a contrast to words that are relevant to the document. This technique was proposed to overcome the problem with Bag of word models

	Kernel	Penality	C (reularization term)
SVM	Linear	l2	1

Table 3
SVM model hyperparameters

	Class weight	Penality	C (reularization term)
Logistic regression	Balanced	l2	10

Table 4
Logistic regression model hyperparameters

and is presented as follows:

$$tf - idf_{t,d} = (1 + \log tf_{t,d}) \cdot \log \frac{N}{df_t}$$

tf: number of occurecnes of t in d
df: number of documents containing t
N: total number of documents

5. Experiments and results

We performed a grid search using three Machine learning classifiers: SVM, multinomial naive bayes, and Logistic regression, to make sure that the hyperparameters were chosen empirically rather than randomly.

We used the default parameters for the Multinomial naive Bayes classifier, as presented in the sickit–learn library. We report the accuracy value according to the number of TF-IDF features. We selected the maximum number of elements after performing grid-search on the TF-IDF vectorizer of Sklearn library[1].

We formed our models again without the empty words list and we obtained the results shown in Figure 6.

from the two figures, we can observe that the accuracy improves with the increase in the number of tf-idf features, and the removal of stop_words increased the accuracy by +0.6%.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html

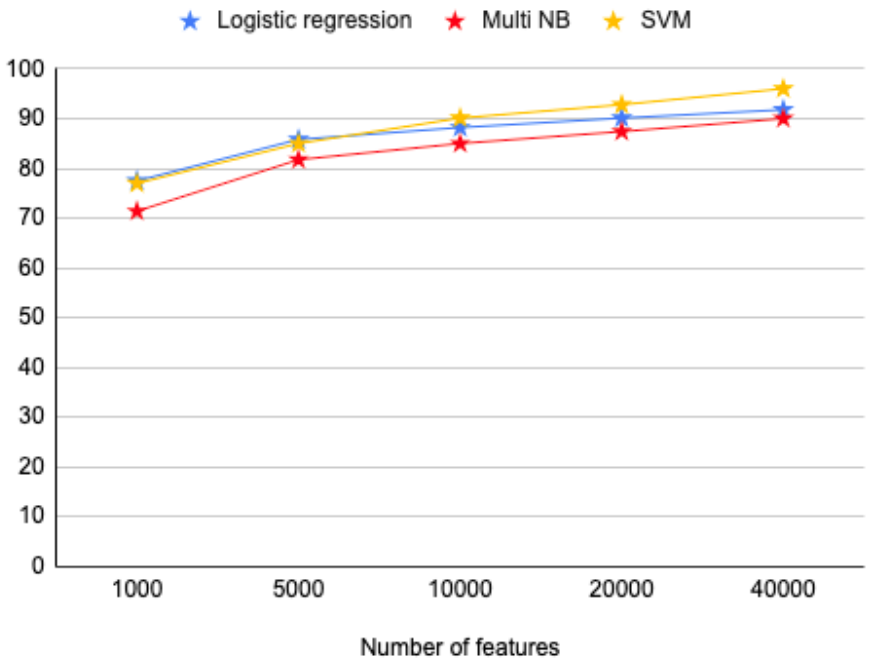

Figure 6: Accuracy level according to the number of features used in TF-IDF without stopwords.

6. Conclusion

In this paper, we have seen the problem with the identification of Maghrebi dialects used in social media, where we presented a large dataset. We demonstrate the effectiveness of machine learning approaches to distinguish between the Algerian, Tunisian, and Moroccan dialects. The problem with TF-IDF is that it cannot represent nor encode the similarity between words in the document since each word is independently presented as an index. Hence Word embedding are an excellent alternative.

For future work, we aim to explore other NLP tasks using this data with word embedding features, such as sentiment analysis, offensive language detection, and translation.

References

[1] I. Guellil, H. Saâdane, F. Azouaou, B. Gueni, D. Nouvel, Arabic natural language processing: An overview, Journal of King Saud University-Computer and Information Sciences (2019).

[2] A. Farghaly, K. Shaalan, Arabic natural language processing: Challenges and solutions, ACM Transactions on Asian Language Information Processing (TALIP) 8 (2009) 1–22.

[3] R. Al-Sabbagh, R. Girju, Yadac: Yet another dialectal arabic corpus., in: LREC, 2012, pp. 2882–2889.

[4] MustGo , about world languages, arabic (levantine), https://www.mustgo.com/worldlanguages/arabic-eastern/, 2020. Accessed: 2020-07-27.
[5] statcounter , social media stats algeria, https://gs.statcounter.com/social-media-stats/all/algeria, 2020. Accessed: 2020-07-27.
[6] statcounter , social media stats tunisia, https://gs.statcounter.com/social-media-stats/all/tunisia, 2020. Accessed: 2020-07-27.
[7] statcounter , social media stats morocco, https://gs.statcounter.com/social-media-stats/all/Morocco, 2020. Accessed: 2020-07-27.
[8] Qatar Foundation International , infographic: Dialects of the arab world, https://www.qfi.org/blog/infographic-dialects-arab-world/, 2020. Accessed: 2020-07-28.
[9] T. Tobaili, Arabizi identification in twitter data, in: Proceedings of the ACL 2016 Student Research Workshop, 2016, pp. 51–57.
[10] K. Sayadi, M. Liwicki, R. Ingold, M. Bui, Tunisian dialect and modern standard arabic dataset for sentiment analysis: Tunisian election context, in: Second International Conference on Arabic Computational Linguistics, ACLING, 2016, pp. 35–53.
[11] T. Tobaili, Arabizi identification in twitter data, in: Proceedings of the ACL 2016 Student Research Workshop, 2016, pp. 51–57.
[12] I. Guellil, F. Azouaou, Arabic dialect identification with an unsupervised learning (based on a lexicon). application case: Algerian dialect, in: 2016 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), IEEE, 2016, pp. 724–731.
[13] D. Seddah, F. Essaidi, A. Fethi, M. Futeral, B. Muller, P. J. O. Suárez, B. Sagot, A. Srivastava, Building a user-generated content north-african arabizi treebank: Tackling hell, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 1139–1150.
[14] K. Darwish, Arabizi detection and conversion to arabic, arXiv preprint arXiv:1306.6755 (2013).
[15] O. F. Zaidan, C. Callison-Burch, Arabic dialect identification, Computational Linguistics 40 (2014) 171–202.
[16] R. Cotterell, C. Callison-Burch, A multi-dialect, multi-genre corpus of informal written arabic., in: LREC, 2014, pp. 241–245.