International Journal of Rough Sets and Data Analysis
Volume 6 • Issue 3 • July-September 2019

An Arabic Dialects Dictionary Using Word Embeddings

Azroumahli Chaimae, National School of Applied Sciences, Abdel Malek Essaâdi University, Morocco

Yacine El Younoussi, National School of Applied Sciences, Abdel Malek Essaâdi University, Morocco

Otman Moussaoui, National School of Applied Sciences, Abdel Malek Essaâdi University, Morocco

Youssra Zahidi, National School of Applied Sciences, Abdel Malek Essaâdi University, Morocco

ABSTRACT

The dialectical Arabic and the Modern Standard Arabic lacks sufficient standardized language resources to enable the tasks of Arabic language processing, despite it being an active research area. This work addresses this issue by firstly highlighting the steps and the issues related to building a multi Arabic dialect corpus using web data from blogs and social media platforms (i.e. Facebook, Twitter, etc.). This is to create a vectorized dictionary for the crawled data using the word Embeddings. In other terms, the goal of this article is to build an updated multi-dialect data set, and then, to extract an annotated corpus from it.

KEYWORDS

Arabic Dialects, Arabic Natural Language Processing, Microblogs, Word Embeddings, Word2Vec

INTRODUCTION

The Arabic speech community exhibits a phenomenon known as the DIGLOSSIA situation, where different varieties of the same language are used by the same community, and each one of these varieties is used for a specific purpose (Farghaly & Shaalan, 2009). There are three main Arabic varieties: The classical Arabic (CA), the Modern Standard Arabic (MSA), and Arabic colloquial dialects. Classical Arabic is the language usually used in religious and literature contexts. Classical Arabic is fully structured and vowelized (Algahtani, 2011). The modern standard Arabic is the official language used in education, media and formal communications across the different Arabic speaking countries, it is based on the CA's syntax and morphology, but it tends to have a more modern vocabulary and "loanwords" (Alotaibi, 2015). And finally, the Arabic colloquial dialects (AD), or the language used in daily informal conversations, it has no orthographic standards so one word can be written in different forms (Shoufan & Al-Ameri, 2015), additionally, it varies from a region to another across the Arabic countries.

The majority of the Arabic dialects occurred as spoken rather than written. However, and since Social media platforms spread off, very large portions of Arabic web users tend to express themselves in written Arabic dialect format (Al-Azani & El-Alfy, 2017). Therefore, and since it is considered the most used Arabic language format in social Web environments, Arabic dialect needs more textual language resources dedicated to a semantic-level analysis. These resources should contain a standardized format of both MSA and the different dialects of the Arab world. The resulting corpora need to be annotated for machine learning, to handle Arabic ambiguities that can result from

DOI: 10.4018/IJRSDA.2019070102

the omission of diacritics (vowels) or the free word-order nature of Arabic sentence. Besides, the researchers noticed that new vocabulary is created occasionally in the colloquial dialects, that leads to the need for a resource that gets updated automatically while analyzing the new entries.

The goal of this paper is to build an updated multi dialect data set, and then, to extract an annotated corpus from it. Finally, the authors plan to develop an approach that semantically analyzes the newly created vocabulary. This corpus will be semantically analyzed using a distributional representation of words known as the word Embedding. This representation will help greatly in overcoming the challenges of normalization between the different Arabic dialects and MSA. From the perspective of both language learning and NLP, the authors expect that there is much to be gained by developing a multi-dialect orthographic convention that makes use, as much as possible, of the common-core etymological bases that all dialects share with MSA.

The rest of this paper is structured as follows: Section 1 gives background information on the Arabic language and the motivation behind creating a multi-dialect corpus. Section 2 highlights some of the related works on creating Arabic dialectical resources. Section 3 and section 4 illustrates the details of collecting and constructing the Arabic multi dialects data set. The authors conclude by introducing their approach and perspective for future works.

THE NEED OF A MULTI-DIALECT CORPUS FOR ARABIC LANGUAGE

Arabic is considered one of the most used Semitic languages with almost 422 million speakers around 22 countries. In addition, it has a huge sphere of influence in the rest of the world since it is the language of the Quran, the holy book of Islam and was the language of science and technologies in the middle ages (Darwish, 2014; Boudad, Faizi, Oulad Haj Thami, & Chiheb, 2017). Further, the Arabic Language is ranked as the seventh top language, and the fastest growing language on the web as Table 1 shows, with over 140 million internet users in the Middle East and North African countries according to (Miniwatts Marketing Group, 2018) which explain the considerable interest that the Arabic language is gaining from the NLP research community. In this paragraph, the authors will give a brief description of the Arabic script, Arabic morphological complexities and the different varieties of the language.

Arabic Script

The use of Arabic script is noticeably growing in the Arabic script-based web content (Miniwatts Marketing Group, 2015). However, it differs greatly from the Latin script, it is written from right to left, using 28 letters that represent consonants. Some of these letters are similar in shape as they differ only by the number or the position of dots, and the majority of the Arabic letters are written connected to neighbouring letters, which results in them changing their shape according to their position in the word (Azroumahli, El Younoussi, & Achbal, 2018). In addition to these characteristics, Arabic Script lacks the orthographic feature of capitalization, and most of the Arabic script content is written without the use of diacritics that plays the role of altering the pronunciation of phoneme or to distinguish between words of similar spelling (Aabed, Awaideh, Elshafei, & Gutub, 2007).

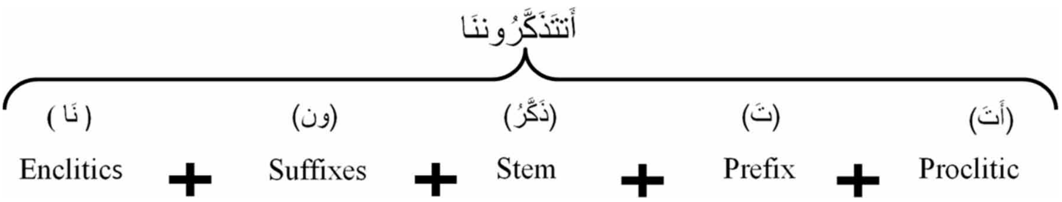
Table 1. Internet usage and population statistics for the Middle East in 2018

	Middle East	Rest of the World
Population	254,438,981 (3.3% of the world)	7,380,318,951
Internet users	147,117,259 (3.6% of the internet users in the world)	3,903,130,324

Arabic Morphology

In NLP research field, morphology is necessary. It represents the identification, analysis and description of the different language unit's structures, that are used to render a much larger set of meaning variations in natural languages such as Arabic (Saad & Ashour, 2010; Soricut & Och, 2015). The Arabic language is widely known for its rich morphology because it is highly inflectional. In term of morphological forms, one word can be structured in a template[1] using the stem[2] of the word, proclitic, prefixes, suffixes and enclitics. Figure 1 demonstrates an example of this template (Will you remember us / أَتتَذَكَّرُوننَا) (Azroumahli et al., 2018). Moreover, Arabic dialects introduce additional clitics that serve new functionality such as negation, mood inflexions, and merged feminine and masculine plurals.

Figure 1. Arabic word structure



Arabic Language Forms: MSA and Dialects

The official forms of Arabic language often used on the internet are either the MSA or the different Arabic Dialects. There are30 major regional Arabic dialects. Arabic dialects and MSA can differ in morphology, lexical properties, phonology and syntax (Graeme, Eduard, & Mark, 2014). These discrepancies are considered to be significant, but the divergence between the dialects themselves can often be of more significance and more disorienting to native speakers of the respective dialects (Graff & Maamouri, 2012), especially, since Arabic dialects have no orthographic standard

Arabic Dialects Challenges

Arabic dialects are considered more challenging than MSA, namely because unlike MSA and CA, it lacks a well-defined spelling system as shown in Table 2. leading to a high sparsity and ambiguity in dialectical Arabic NLP systems (Diab, 2009). Arabic's optional use of diacritics and the inconsistent spelling of some letters lead to a high degree of ambiguity as well. Besides, Arabic dialects maintain some of MSA's complex syntactic phenomena such as irrational plural agreement.

Table 2. A person's name in CA and MSA expressed by more than one variety in dialectical Arabic

MSA & CA	Moroccan	Egyptian	Syrian
عبد القادر Abd Al-Kader	عبدلقادر Abdlkader	عبد الجادر Abd El-Gader	عبد الآدر Abd El-Aader

A Web Based Multi Dialect Corpus

One of the key challenges in computational Arabic web content processing, particularly semantic level analysis and synthesis, is how to handle the morphological and the syntactic differences between the variant Arabic dialects and MSA (Graeme et al., 2014). Besides, the Arabic NLP community suffers greatly from the lack of written Arabic dialects resources, so, there is a lot to gain by developing a corpus that assembles, as much as possible, the common core etymological bases that all dialects share with the MSA. The main objective of this work is to build an Arabic multi dialect MSA corpus by taking advantage of the massive growth of Arabic content on the web, especially social media environments, and the fact that both MSA and Arabic dialects share the same Arabic script except for using Arabizi[3]. Therefore, in theory, the researchers can use the same approach to process both MSA and the different Arabic dialects varieties of the same corpus. In the following sections, the authors will site some of the related works on creating Arabic dialectical resources, and then proceed by describing their data design.

Related Works

The research interest in creating annotated Arabic dialects resources for computational language processing has emerged in recent years. In the following section, the authors will site some of the existing Arabic dialects corpora and data sets and their building description, while highlighting the approaches and methods followed for each.

In (Sansò, Linguistica, Pavia, & Nuova, 2005), the authors suggested an approach dedicated to the storage of Mediterranean language data for quantitative and qualitative typological analysis as a part of the Med TYP project. During their analysis, they debated that the Mediterranean dialects could not be identified as a linguistic area in the traditional sense.

In (Graff & Maamouri, 2012), the authors presented the approach they followed to update the dictionaries they were developed by the Linguistic Data Consortium at Georgetown University Press to three bilingual dictionaries that contain Moroccan, Syrian and Iraqi dialects for English speakers. Their goal was to standardize a uniform orthographic strategy across Arabic dialects while applying their LMF[4]-XML structure to the lexicon of the data sources and IPA[5] spellings. For their new data design, they used the Arabic script and International Phonetic Alphabet orthographies.

In (Boujelbane, Khemakhem, Ayed, & Belguith, 2013), the authors described the creation of a bilingual dictionary which contains the Tunisian dialect and MSA to create a language model for speech recognition system dedicated to the Tunisian Broadcast News. They developed a tool called Tunisian dialect Translator that enables the production of Tunisian dialect texts and the enrichment of the MSA-Tunisian dialect dictionary.

The authors of (Al-sabbagh & Girju, 2012), presented their first phase on building a multi-genre and multi dialectal Arabic corpus YArabic DialectAC for Egyptian dialect. It incorporates data from multiple web sources including microblogs, online knowledge market services and blogs. Their corpus offers linguistic analyses such as the POS tagging and base phrase chunking.

In (Lulu & Elnagar, 2018), a data set of Egyptian (EGP), Levantine (LEV), and Gulf –including Iraqi (GLF) dialect was presented. The resulted corpus contained 33k sentences. This corpus was used to implement a dialect identification system using various deep neural network models.

Other interesting works on building multidialectal Arabic text corpora using the web resources are the works described in (Almeman & Lee, 2013) and (Al-Twairesh, Al-Khalifa, Al-Salman, & Al-Ohali, 2017). The (Almeman & Lee, 2013) authors' approach was to survey and collect a specific dialect text corpus, these dialects were categorized into four main origins, Gulf, Levantine, Egyptian and North African. This categorization was conducted using accent identification. As for (Al-Twairesh et al., 2017) authors, their goal was to build a corpus of Arabic tweets annotated for sentiment analysis, their corpus consists of tweets written in MSA and the Saudi dialect and contains 17,573 tweets.

Recently, more works have been emerged on creating Arabic dialects corpora dedicated to sentiment classification. The works presented in (Tartir & Abdul-Nabi, 2017) and (Baly, Alaa Khaddaj, Hazem Hajj, & Wassim El-Hajj, 2018) are examples of these corpora. The authors of (Tartir & Abdul-Nabi, 2017) proposed a semantic approach to discover user attitudes and business insights from Arabic dialect texts that were generated from social media sources. They applied their approach on a corpus that they collected using Tweet Archivist from posts relevant to certain topics and brand names. The work described in (Baly & Alaa Khaddaj, Hazem Hajj, Wassim El-Hajj, 2018) presented an Arabic Sentiment Twitter Dataset for the Levantine dialect. Their corpus contained 4,000 tweets annotated as follow: the overall sentiment of the tweet, the target to which the sentiment was expressed, how the sentiment was expressed, and the topic of the tweet. The annotations that they used improved the performance of a baseline sentiment classifier.

CREATING THE CORPUS: DATA DESIGN

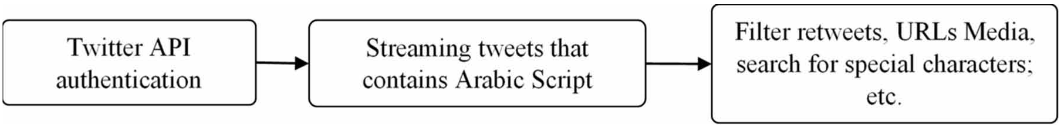
Due to the rise of modern communication technologies and social networking sites such as Instagram, Facebook and Twitter, massive amounts of text data are being generated in real-time. Moreover, these social web environments are considered a powerful tool for disseminating information, and a rich resource for opinionated text since it contains various opinions on many different topics: politics, business, economic, social life etc. As a result, Social media sites make our corpus not only usable for information extraction, but also for answering computer-human interaction questions, such as how web users communicate to express opinions, show sentiments and take sides in arguments.

In this section, the authors will present the details of collecting and constructing their data set of Arabic tweets and Facebook /Instagram comments that contains both Arabic dialects and MSA, since, as (Al-sabbagh & Girju, 2012) explained, the language used in social media is known to be highly dialectal.

Arabic Language Forms: MSA and Dialects

Twitter and Facebook APIs make it possible to steam written data to enable a programmatic analysis of tweets and Facebook comments. The methodology of creating a corpus using Twitter API involves three main steps as Figure 2 demonstrates; Preparing the python authentication using Tweepy[6], streaming and filtering the Arabic tweets from Arabic users using function-based harvesting, where only the live tweets that contain Arabic script were extracted. The same methodology was followed to collect Facebook comments using Facebook Excel Add In[7]; once the Data connection was configured, they specified the Facebook pages and groups that will be populated with live Facebook data.

Figure 2. The steps of creating a corpus using Twitter API


Twitter API authentication
Streaming tweets that contains Arabic Script
Filter retweets, URLs Media, search for special characters; etc.

To determine the ratio of Facebook and Twitter live data that the authors should collect, they conducted a study using Google Trends[8] tool. The results are shown in Figure 3 and Figure 4 illustrates that contrariwise to the middle east web users who tend to utilize both these social media platforms, the North African prefer using Facebook much more than Twitter. Thus, and to collect all the verities of Arabic dialects, the authors focused on the middle eastern dialects using Twitter, and on the North African dialects using Facebook. The statistics of this study can be accessed from [9]https://github.com/AzChaimae/Arabic-Facebbok-and-Twitter-users-2018/tree/master10.

Figure 3. Active social media North African users for the last 12 months (09-01-2019)

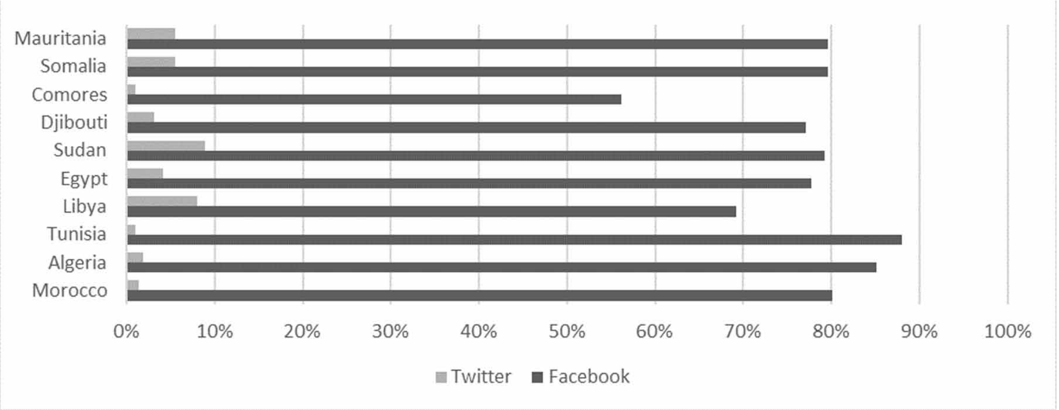
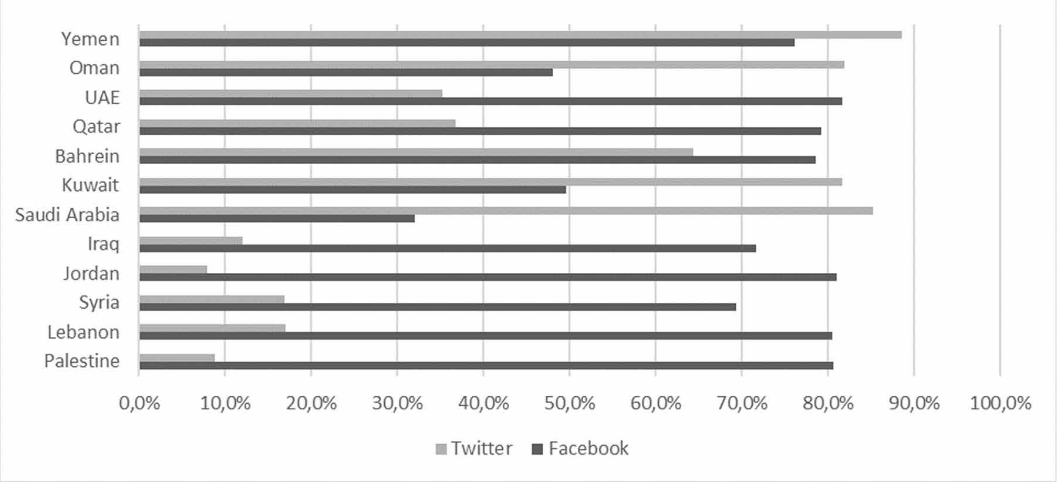


Figure 4. Active social media Middle Eastern users for the last 12 months (09-01-2019)



Data Set Cleaning and Preprocessing

Web data generated from microblogs is known to be noisy, therefore our crawled data should be preprocessed and normalized to get high-quality data (Ben Abdessalem Karaa & Dey, 2017) (Singh, Dey, Ashour, & Santhi, 2017). The crawled data is cleaned from non-Arabic words that could be URLs, username mentions (@user), or any non-Arabic character's expression. The other preprocessing steps involve the word tokenization and word Normalization. The tokenization is used to reduce the sparsity of observed forms by using NLTK[11] library. The word Normalization step included uniforming the shape of some Arabic letters and removing diacritics to make text in a consistent form as suggested by (S. Alotaibi & Khan, 2017). Table 3 summarize the normalization cases followed.

The last step was to remove all the Arabic stop words based on Mohamed Taher's list[12], and the tweets that contain only one word since the word representation approach that the authors intended to use depends on the word's context, and finally remove any remaining duplicated data. Table 4 shows the output words resulted from this web crawling and cleaning process.

Table 3. Normalization cases

Letters	Normalization
{/ $\overline{a}lif$ / ا, /Aa/ آ, /hamza -a-/ أ, /hamza -i-/ إ, /ū/ ٱ }	{/âlif/ ا}
{/tā/ ة, /hā/ه }	{/hā / ه}
{/ yā / ي, /yeh/ ى}	{/ yeh / ى}

Table 4. Statistics of the collected Data

	Collected entries	Cleaned & filtered entries	Number of words
Example	@Fsleyt: 🤓: وش تخصصك؟	وش تخصصك	[‘وش’,‘ تخصصك‘]
Tweets			
Facebook comments			

ARABIC DIALECTS' WORD EMBEDDING

In order to process all the variants of Arabic dialect, the researchers needed an NLP approach that specifies the Arabic language variant beforehand. Nonetheless, the authors can overcome this step by benefiting from neural network methods and use vectorized distributed word representation known by the name of word embeddings. This representation is famous for its capability of capturing morphological, syntactic and semantic information about words, by enabling efficient computation of word similarities (Azroumahli et al., 2018). The authors used this approach to create an Arabic dialects vectorized dictionary, aiming for the resolving of the no orthographic standard problem and normalizing Arabic dialects.

word embeddings are low dimensional, dense real value vectors. These vectors are built by considering as input a word and its context in a large corpus, this corpus leads to the learning of the vectors presenting the output vocabulary. There are three main word embedding's learning models in the literature. 1) The global matrix factorization models such as the latent semantic analysis (LSA) (Dhillon, Foster, & Ungar, 2011). 2) the local context window models, such as the Wor2Vec model (Mikolov, Chen, Corrado, & Dean, 2013), 3) and the word occurrence in a corpus model known as the GloVe Model (Pennington, Socher, & Manning, 2014). To achieve the most accurate results, and based on (Dahou, Xiong, Zhou, & Haddoud, 2016) analysis, The authors opted for the Skip-Gram (SG) neural network architecture available in the Word2Vec[13] tool to create a low dimensionality vectorized vocabulary.

Parameter Settings

The Skip-gram is Word2Vec architecture based on the probabilistic feedforward neural network language model (Neeraj, Yoshua, Réjean, Pascal, & Christian, 2003). Word2Vec architectures use log-linear[14] training to capture semantic information; the algorithm pre-train a single projection matrix $W \in \Re^{d\times|V|}$ where d is the embedding dimension and V is the vocabulary. The embeddings are built by maximizing the likelihood of word prediction of their context and vice versa (Sallam, Mousa, & Hussein, 2016). The Skip-Gram method uses a word as an input in a log-linear classifier with a continuous projection[15] layer and predicts the input word's surroundings in a context within a window size Figure 5 shows the process the authors followed to build their word representations using this architecture.

The hyper parameters used to construct the different embeddings for the Skip-Gram architecture are: The Dimensionality of the vectors[16], the window size of a word's contexts, the minimum word occurrences, and the number of negative samples. After different alteration of the Skip-Gram training parameters, and in order to maximize the evaluation result, the authors opted for the parametrizations shown in Table 5. The negative sampling method was used to improve the quality of the frequent words' vectors quality and to speed up the training process. The resulting vectorized vocabulary contains a 200 dimension vectorized words that preserve the semantic features like the examples shown in Table 6. Further, the vectors can cluster the words into specific categories, these clustered categories represent the word similarities as shown in Figure 6 and Table 7.

Figure 5. The steps of creating word representations for a twitter database

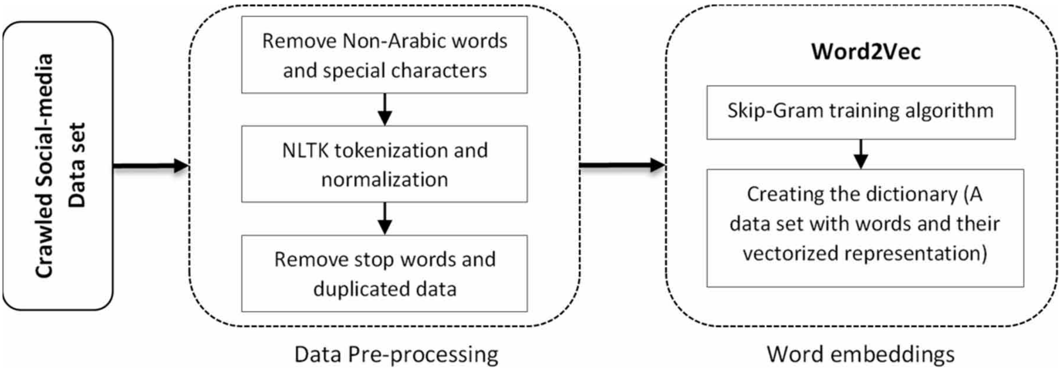

Crawled Social-media Data set
Remove Non-Arabic words and special characters
NLTK tokenization and normalization
Remove stop words and duplicated data
Data Pre-processing
Word2Vec
Skip-Gram training algorithm
Creating the dictionary (A data set with words and their vectorized representation)
Word embeddings

Table 5. Skip-Gram training parameters

Dimensionality	Window size	Sample	Minimum word count	Negative sampling
200	5	$1\times e^{-5}$	10	10

Word Embeddings Evaluation

To evaluate the obtained word embeddings, the authors choose the word analogy test to report the accuracy of the representations. Word Analogy test was first introduced by the authors of (Mikolov et al., 2013), where the goal is to solve analogy questions by finding a term x for a given term y, so that $x:y$ is the best resemblance to a sample relationship $a:b$.

Table 6. Examples of the data set output

Words	200 dimension vectorized representation
/Ab/ أب	$\left[-0.00786473, 0.0026544, \text{.....}, 0.0194475, -0.0288277, 0.0131017\right]$
/Yii/ يي	$\left[-0.0224097, 0.0120036, \text{.....}, -0.0763094, -0.112953, 0.0508491\right]$

Figure 6. A visualization of one of the resulting embeddings using TensorFlow embeddings projector[17]

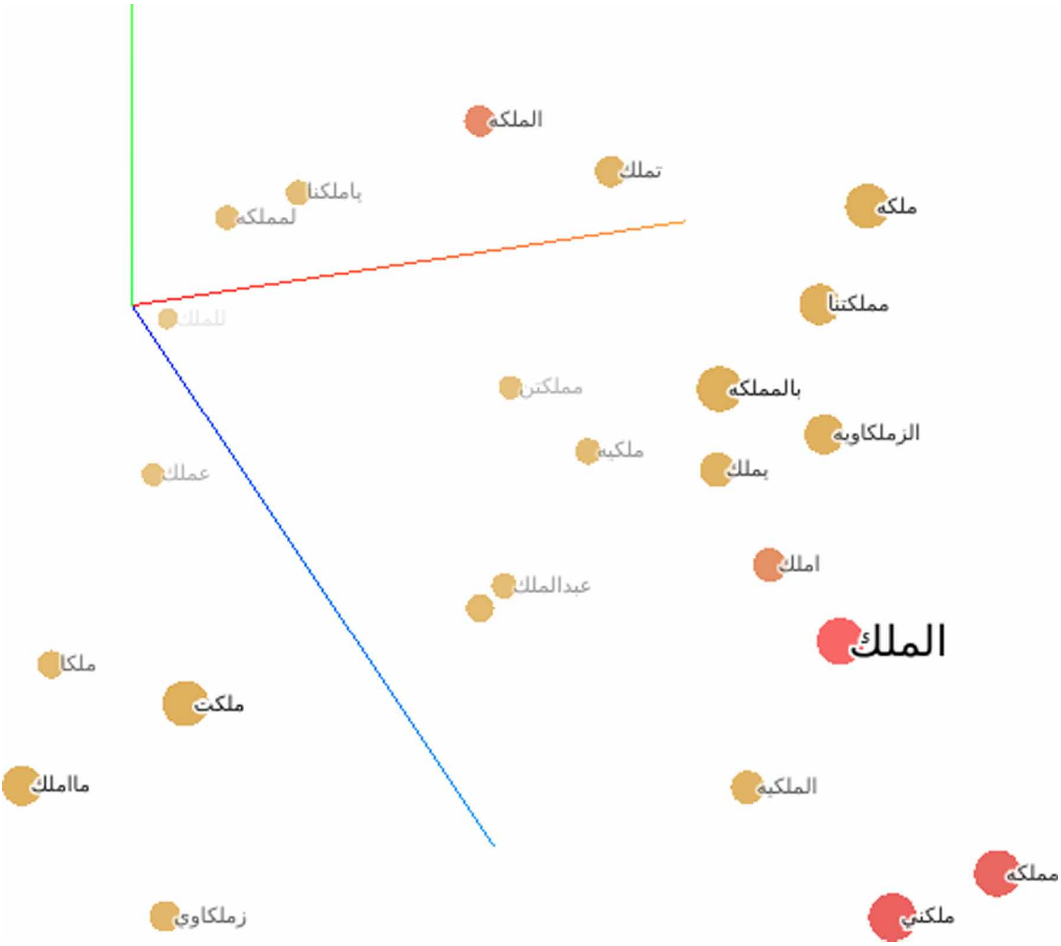


Table 7. An example of a word and its nearest points

Label	الملك (the king)
Examples of nearest points	**Cosine similarity**
الملك (king) المملكه (The kingdom) ملكنا (Our king) عرش (Throne) الدوله (The country) الحكم … (The judgment)	0.251 0.421 0.567 0.693 0.782 0.839

This test was conducted by using an enhanced and translated version of Google's word analogy test benchmarks, aided by the available benchmarks published by the authors of (Baly, Alaa Khaddaj, Hazem Hajj, & Wassim El-Hajj, 2018). The benchmarks contain 71974 relations and cover 11 relation

Table 8. Word analogy relation types

Morphosyntactic	Semantic
Comparative Plurals pairs Verb to noun	Capital cities Common capital cities Currency Family Man Woman Nationality adjectives (male) Opposite
35978	**35996**

types, 4 of them are semantic and 7 are morphosyntactic (see Table 8). An example of a typical analogy relation from our benchmarks test set contains two pairs of words as is illustrated in Table 9.

To compute the vector's analogies, the authors needed to recover the relations between the word vectors. Figure 7 presents the algorithm used to compute these relations. Some of the words that are present in the benchmarks, do not exist in the corpus since they are not often used by social media users. An accuracy of 0.46 was obtained on the resulting embeddings using the corpus with the hyperparameters shown in Table 5. It is lower than those obtained on English and other Latin languages. However, the MSA content in social media corpus is minor compared to the Arabic dialects content, which results in the rarity of MSA words, in addition, the benchmarks used in this evaluation are written in MSA. Thus, the obtained accuracy is considered a good result, due to the morphological complexity of MSA and Arabic dialects. The source code and the corpus described in this paper can be obtained from https://github.com/AzChaimae/Arabic-dialects-Corpus.

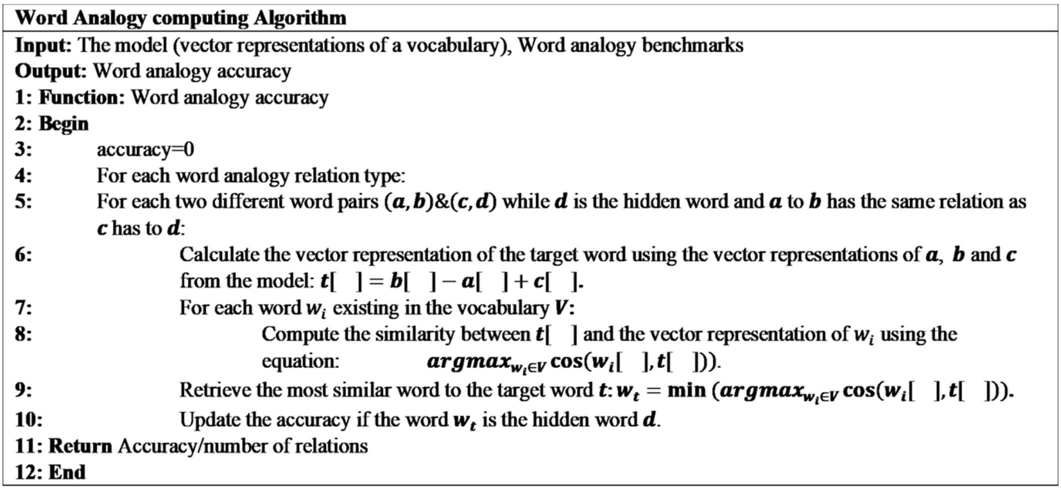
CONCLUSION

In this article, the authors presented the steps of creating an Arabic vectorized dictionary that contains the different variations of Arabic dialects. This work relied on two steps, the first one being the use of social media as a source for crawling the data since it is the most comprehensive source of live, public conversation. The second one is the use of word Embeddings methods which had produced impressive results, by speeding up the training process with the use of Word2Vec's negative sampling approach, and enabling the learning of word representations from a large-scale corpus of words, taking as input raw texts without any other source of information other than the word's context from the raw text itself. The resulting corpus could serve as a solution for the obvious lack of resources in the Arabic dialects processing world since it contains both the MSA and the different variations of Arabic dialects.

Table 9. Word Analogy test-set example

Relation type	Word pair 1		Word pair 2	
Capitals of the world	لشبونة Lisbon	البرتغال Portugal	مدريد Madrid	اسبانيا Spain
Family relation	أبي Dad	أمي mum	عمي My uncle	عمتي My aunt
Opposite relations	نام Slept	استيقظ Woke up	حزين Sad	سعيد Happy

Figure 7. Word analogy computing's algorithm



```
Word Analogy computing Algorithm
Input: The model (vector representations of a vocabulary), Word analogy benchmarks
Output: Word analogy accuracy
Function: Word analogy accuracy
Begin
    accuracy=0
    For each word analogy relation type:
    For each two different word pairs $(a, b)\&(c, d)$ while $d$ is the hidden word and $a$ to $b$ has the same relation as $c$ has to $d$:
        Calculate the vector representation of the target word using the vector representations of $a$, $b$ and $c$ from the model: $t[\ ] = b[\ ] - a[\ ] + c[\ ]$.
        For each word $w_i$ existing in the vocabulary $V$:
            Compute the similarity between $t[\ ]$ and the vector representation of $w_i$ using the equation: $argmax_{w_i \in V} \cos(w_i[\ ], t[\ ])$.
        Retrieve the most similar word to the target word $t$: $w_t = \min(argmax_{w_i \in V} \cos(w_i[\ ], t[\ ]))$.
        Update the accuracy if the word $w_t$ is the hidden word $d$.
Return Accuracy/number of relations
End
```

From an implementational perspective, the aboutness task involves experimenting with different other sources for creating other corpora, using the obtained model to build several binary classifiers to detect subjectivity and sentiments in both the modern standard Arabic and dialects. Finally, adding Arabizi words to the corpus since they are widely used by Arabic users on the web especially in north African countries.

REFERENCES

Aabed, M. A., Awaideh, S. M., Elshafei, A. R. M., & Gutub, A. A. (2007). Arabic diacritics based steganography. In *ICSPC 2007 Proceedings - 2007 IEEE International Conference on Signal Processing and Communications* (pp. 756–759). IEEE Press. doi:10.1109/ICSPC.2007.4728429

Al-Azani, S., & El-Alfy, E. S. M. (2017). Using Word Embedding and Ensemble Learning for Highly Imbalanced Data Sentiment Analysis in Short Arabic Text. *Procedia Computer Science*, *109*, 359–366. doi:10.1016/j.procs.2017.05.365

Al-sabbagh, R., & Girju, R. (2012). YADAC: Yet another Dialectal Arabic Corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (pp. 2882–2889). Academic Press.

Al-Twairesh, N., Al-Khalifa, H., Al-Salman, A., & Al-Ohali, Y. (2017). AraSenTi-Tweet: A Corpus for Arabic Sentiment Analysis of Saudi Tweets. *Procedia Computer Science*, *117*, 63–72. doi:10.1016/j.procs.2017.10.094

Algahtani, S. (2011). *Arabic Named Entity Recognition: A Corpus-Based Study*. University of Manchester.

Almeman, K., & Lee, M. (2013). Automatic building of Arabic multi dialect text corpora by bootstrapping dialect words. In *Proceedings of the 2013 1st International Conference on Communications, Signal Processing and Their Applications ICCSPA 2013*. Academic Press. doi:10.1109/ICCSPA.2013.6487247

Alotaibi, F. S. S. (2015). *Fine-grained Arabic Named Entity Recognition*. University of Birmingham.

Alotaibi, S., & Khan, M. B. (2017). Sentiment Analysis Challenges of Informal Arabic Language. *International Journal of Advanced Computer Science and Applications*, *8*(2), 278–284. doi:10.14569/IJACSA.2017.080237

Azroumahli, C., El Younoussi, Y., & Achbal, F. (2017, December). An overview of a distributional word representation for an Arabic named entity recognition system. In Proceedings of the International Conference on Soft Computing and Pattern Recognition (pp. 130-140). Springer. doi:10.1007/978-3-319-76357-6_13

Baly, R., Khaddaj, A., Hajj, H., El-Hajj, W., & Shaban, K. B. (2019). ArSentD-LEV: A multi-topic corpus for target-based sentiment analysis in Arabic levantine tweets.

Ben Abdessalem Karaa, W., & Dey, N. (2017). *Mining Multimedia Documents* (1st ed.). Chapman and Hall/CRC.

Boudad, N., Faizi, R., Oulad Haj Thami, R., & Chiheb, R. (2017). Sentiment analysis in Arabic: A review of the literature. *Ain Shams Engineering Journal*. doi:10.1016/j.asej.2017.04.007

Boujelbane, R., Khemakhem, M. E., Ben Ayed, S., & Belguith, L. H. (2013). Building bilingual lexicon to create Dialect Tunisian corpora and adapt language model. In *Proceedings of the Second Workshop on Hybrid Approaches to Translation* (pp. 88–93). Association for Computational Linguistics.

Dahou, A., Xiong, S., Zhou, J., & Haddoud, M. H. (2016). Word Embeddings and Convolutional Neural Network for Arabic Sentiment Classification. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING-16)* (pp. 2418–2427). Academic Press.

Darwish, K. (2014). Arabic Information Retrieval. *Foundations and Trends in Information Retrieval, 7*(4), 239–342. doi:10.1561/1500000031

Dhillon, P. S., Foster, D., & Ungar, L. (2011). Multi-view learning of word embeddings via CCA. In *Advances in Neural Information Processing System (NIPS 2011)*. Academic Press. Retrieved from http://papers.nips.cc/paper/4193-multi-view-learning-of-word-embeddings-via-cca

Diab, M. (2009). Second Generation AMIRA Tools for Arabic Processing: Fast and Robust Tokenization, POS tagging, and Base Phrase Chunking. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools* (pp. 285–288). Academic Press. Retrieved from http://www.elda.org/medar-conference/pdf/56.pdf

Farghaly, A., & Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing*, *8*(4), 1–22. doi:10.1145/1644879.1644881

Graeme, H., Eduard, H., & Mark, J. (2014). Natural Language Processing of Semitic Languages. In Theory and Applications of Natural Language Processing. Springer. doi:10.1007/978-3-642-45358-8

Graff, D., & Maamouri, M. (2012). Developing LMF-XML Bilingual Dictionaries for Colloquial Arabic Dialects. In *Lrec 2012 - Eighth International Conference on Language Resources and Evaluation* (pp. 269–274). Academic Press.

Lulu, L., & Elnagar, A. (2018). Automatic Arabic Dialect Classification Using Deep Learning Models. *Procedia Computer Science*, *142*, 262–269. doi:10.1016/j.procs.2018.10.489

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the IJCAI International Joint Conference on Artificial Intelligence* (pp. 4069–4076). Academic Press. doi:10.1162/153244303322533223

Miniwatts Marketing Group. (2015). Internet Users in 2015 in the Middle East and the World. Retrieved from http://www.internetworldstats.com/stats5.htm#me

Miniwatts Marketing Group. (2018). Middle East Internet Statistics, Population, Facebook and Telecommunications Reports. Retrieved from https://www.internetworldstats.com/stats5.htm

Neeraj, S. S., Yoshua, B., Réjean, D., Pascal, V., & Christian, J. (2003). A Neural Probabilistic Language Model. *Journal OfMachine Learning Research*, *3*, 1137–1155. doi:10.1162/153244303322533223

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Academic Press. doi:10.3115/v1/D14-1162

Saad, M., & Ashour, W. (2010). Arabic Morphological Tools for Text Mining. In *Proceedings of the 6th International Conference on Electrical and Computer Systems (EECS'10)*. Academic Press.

Sallam, R. M., Mousa, H. M., & Hussein, M. (2016). Improving Arabic Text Categorization using Normalization and Stemming Techniques. *International Journal of Computers and Applications*, *135*(2), 38–43. doi:10.5120/ijca2016908328

Sansò, A., Linguistica, D., Pavia, U., & Nuova, S. (2005). MED-TYP: A Typological Database for Mediterranean Languages Introduction : the MED-TYP project. *Architecture*, 1157–1160.

Shoufan, A., & Al-Ameri, S. (2015). Natural Language Processing for Dialectical Arabic. *Survey*, 36–48. Retrieved from http://www.aclweb.org/anthology/W15-3205

Singh, A., Dey, N., Ashour, A. S., & Santhi, V. (Eds.). (2017). *Web semantics for textual and visual information retrieval*. Hershey, PA: IGI Global; doi:10.4018/978-1-5225-2483-0

Soricut, R., & Och, F. J. (2015). Unsupervised Morphology Induction Using Word Embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT 2015)* (pp. 1626–1636). Academic Press.

Tartir, S., & Abdul-Nabi, I. (2017). Semantic Sentiment Analysis in Arabic Social Media. *Journal of King Saud University - Computer and Information Sciences, 29*(2), 229–233. doi:10.1016/j.jksuci.2016.11.011

ENDNOTES

[1] An abstract pattern in which roots are inserted. The template like the root provides some meaning component.

[2] Or the root of a word: a sequence of mostly three consonantal letters which together signify some abstract meaning. The Arabic Verbs and nouns are derived from the stems.

[3] Arabizi, Arabish or Arab-France is the Arabic written using Latin characters in transliterated mode and numerals to represent Arabic letters with no phonetic equivalent in English or French (Darwish, 2014).

[4] Lexical Markup Framework, ISO 24613

[5] International Phonetic Alphabet, it is used for characters pronunciation

[6] An open source python Library used mainly to access the Twitter API.

[7] Facebook Excel Add-In is a tool that sets a connection with live Facebook data directly from Microsoft Excel, https://download.cnet.com/Excel-Add-In-for-Facebook/3000-2065_4-76476610.html

[8] https://trends.google.fr/trends/?geo=FR

[9] The North African and the middle eastern Facebook and twitter users' statistics will be released before April 2019.

[10] The North African and the middle eastern Facebook and twitter users' statistics will be released before April 2019.

[11] Natural language Toolkit for building Python programs to work with human language data.

[12] The MIT License 2016 available at https://github.com/mohataher/arabic-stop-words/blob/master/list.txt

[13] http://code.google.com/p/word2vec/

[14] Log-linear classifiers, also known as Maximum Entropy classifiers, produce a probability distribution by incorporating linguistically important features and allowing the automatic building of language-independent, retargetable NLP modules.

[15] The projection layer maps the word indices of an n-gram context to a continuous vector space. They are used to reduce the dimensionality of representation without reducing its resolution.

[16] The size of the Neural Network layers

[17] http://projector.tensorflow.org/