An Algerian Corpus and an Annotation Platform for Opinion and Emotion Analysis

Leila Moudjari, Karima Akli-Astouati, Farah Benamara

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1202–1210
Marseille, 11–16 May 2020

An Algerian Corpus and an Annotation Platform for Opinion and Emotion Analysis

Leila Moudjari, Karima Akli-Astouati, Farah Benamara
RIIMA Laboratory USTHB Bab Ezzouar 16111, Algiers Algeria
IRIT Université de Toulouse, 118 Route de Narbonne,
l.moudj11@gmail.com/lmoudjari@usthb.dz, kakli@usthb.dz, farah.benamara@irit.fr

Abstract

In this paper, we address the lack of resources for opinion and emotion analysis related to North African dialects, targeting Algerian dialect. We present TWIFIL (TWItter proFILing) a collaborative annotation platform for crowdsourcing annotation of tweets at different levels of granularity. The plateform allowed the creation of the largest Algerian dialect dataset annotated for both sentiment (9,000 tweets), emotion (about 5,000 tweets) and extra-linguistic information including author profiling (age and gender). The annotation resulted also in the creation of the largest Algerien dialect subjectivity lexicon of about 9,000 entries which can constitute a valuable resources for the development of future NLP applications for Algerian dialect. To test the validity of the dataset, a set of deep learning experiments were conducted to classify a given tweet as positive, negative or neutral. We discuss our results and provide an error analysis to better identify classification errors.

Keywords: Crowdsourcing annotation platform, Algerian dialect, sentiment analysis, emotion detection, subjectivity lexicon

1. Introduction

Currently, there are more than 4 billion Internet users worldwide. More than 50% of the North African population has access to the Internet. The region has also seen a growth of more than 17% in the number of social media users compared to 2017 [1]. In Algeria, more than 50% of the population are registered users on different social platforms and around 46% of them use mobile devises for such activity [2]. These numbers represent a growth of 17% in social media use and more than a 19% growth in the use of mobile devises for such platforms. Twitter users in Algeria reached 8.73% in August 2019 compared to August 2018 (2.96%). The number has almost tripled over a year, making Twitter the third most used platform by active social media users [3]. On social media platforms, 76% of users express their sentiments by clicking corresponding buttons when available, such as "Like", "Dislike". Around 50% expresses views or sentiments using "emoticons", "emojis" or "smileys". Across the Arab region, more than 30% of the users use Arabic script and 26% uses Latin script (mostly English and French) and about 15% combine both (Salem, Feb 5 2017). Compared to other Arabic dialects, the North African dialects have other peculiarities, as several languages are used in everyday conversations. For example, the expression "*Nro7o ensemble?*" is a combination of the French word "*ensemble*" meaning "*together*", and the Arabic word "*nro7o (نرُوحُو)*", meaning "*we go together*".

Sentiment analysis and emotion detection in Arabic have been widely studied (Baly et al., 2017; Al-Smadi et al., 2018; Abo et al., 2018). Most related work focus on Modern Standard Arabic (MSA), although a few investigated Arab dialects, such as Jordanian (Atoum and Nouman, 2019; Duwairi, 2015), Egyptian (Shoukry and Rafea, 2012), Iraqi (Alnawas and Arici, 2019), Levantine (Baly et al., 2019; Qwaider et al., 2019) and Tunisian (Medhaffar et al., 2017). North African dialects, including Algerian dialects (ALGD) are less normalised compared to MSA. They have been enriched by many languages over the years, which resulted in a complex linguistic situation. Also, we found a significant lack of resources for most of these dialects such as lexicons, dictionaries, and annotated corpora. In this paper, we address the lack of resources for opinion and emotion analysis related to North African dialects, targeting Algerian dialect. We present TWIFIL (TWItter proFILing) a collaborative annotation platform for crowdsourcing annotation of tweets at different levels of granularity. The platform allowed the creation of the largest Algerian dialect dataset annotated for both sentiment (9,000 tweets), emotion (about 5,000 tweets) and extra-linguistic information including author profiling (age and gender). The annotation resulted also in the creation of the largest Algerien dialect subjectivity lexicon of about 9,000 entries which can constitute a valuable resources for the development of future NLP applications for Algerian dialect.

This paper is organised as follows: Section 2 provides a general overview of opinion and emotion analysis (OEA) in ALDG. Section 3 introduces the specificities of the ALGD. The annotation platform is described in Section 4 and experiments in Section 5. We finally conclude providing some perspectives for future work.

2. Related Work

Over the years OEA has been widely used in a variety of applications such as marketing and politics, etc. These have inspired several methods ranging from lexicon-based approaches (Al-Moslmi et al., 2018) to corpus-based (Abdul-Mageed and Diab, 2012) to recently Deep learning (Al-Smadi et al., 2018).

As mentioned in the introduction, numerous studies on Arabic sentiment analysis have been carried out in recent years (Abdul-Mageed and Diab, 2012; Nabil et al., 2015; Badaro et al., 2018). The Arabic dialects are a variety of MSA which includes languages with less normalisation and standardisation (Saadane and Habash, 2015). They differ from

[1] https://wearesocial.com/blog/2018/01/global-digital-report-2018 (visited on 23rd, November 2019)

[2] https://www.slideshare.net/wearesocial/digital-in-2018-in-northern-africa-86865355

[3] http://gs.statcounter.com/social-media-stats/all/algeria/2019

MSA on all levels of linguistic representation, from phonology and morphology to lexicon and syntax.

It is worth mentioning that the highest proportion of available resources and research publications in Arabic OEA are devoted to MSA. Regarding Arabic dialects, the Middle-Eastern and Egyptian dialects received the largest share of all research effort and funding. On the other hand, very little work has been conducted for the OEA of the Maghrebian dialects (Medhaffar et al., 2017). In addition, research into ALGD is rare which resulted in a lack of resources.

The proposed Arabic OEA approaches focus mainly on MSA where few of Arabic dialects have been explored, Jordanian (Atoum and Nouman, 2019; Duwairi, 2015), Egyptian (Shoukry and Rafea, 2012), Iraqi (Alnawas and Arici, 2019), Levantine (Baly et al., 2019; Qwaider et al., 2019) and Tunisian (Medhaffar et al., 2017). Even though, the community is attracting more and more attention to the Arabic dialects with competitions such as the 2018 Semantic Evaluation competition first task[4]. Which included five sub-tasks on inferring the affectual state of a person from their tweet: 1. emotion intensity regression, 2. emotion intensity ordinal classification, 3. valence (sentiment) regression,4. valence ordinal classification, and 5. emotion classification. For each sub-task, labeled data were provided for English, Arabic, and Spanish (Mohammad et al., 2018).

North African countries are known for their diversity in spoken dialects, which in recent years have generated huge volumes of written data on social media, such as Algerian Arabic, which is widely used on social networks.

In (Qwaider et al., 2019) the authors studied the feasibility of using MSA approaches and apply them directly on a Levantine corpus. Results were as expected, they obtained not more than 60% accuracy. However, when they tested different machine learning algorithms they reached an accuracy of 75.2%. The same approach was adopted to tackle the ALGD. Where the methods of OEA applied to ALGD were the same as those applied to MSA. At first it seemed promising, although yielded significantly low performances (Saadane and Habash, 2015). So it was deemed necessary to develop solutions and build resources for the OEA of the ALGD.

(Saadane and Habash, 2015), proposed a list of phonetic rules to be followed, to facilitate the automatic translations of Algerian Arabic and MSA, in both directions. Such tools could be used in several Natural Language Processing (NLP) applications, such as OEA. The authors rely on the CODA spelling model (Conventional Orthography for Dialectal Arabic) proposed by (Habash et al., 2012), for the Egyptian dialect. Furthermore, (Zribi et al., 2014) extend the CODA guidelines to take into account to Tunisian dialect and (Jarrar et al., 2014) have adapted it to the Palestinian dialect.

In the same way, Harrat et al. (2017) present a Maghrebi multi-dialect study including dialects from Algeria, Tunisia and Morocco that they compare to MSA.

Harrat et al. (2014), constructed a parallel dataset for Algerian dialects, with the objective of building Machine Translation solutions for MSA and ALGD, in both directions.

Mataoui et al. (2016), presented a Lexicon-Based Sentiment Analysis Approach for Vernacular Algerian Arabic, the approach addresses specific aspects of the ALGD fully utilised in social networks. A manually annotated corpus and three lexicons, (negation words lexicon, intensification-words Lexicon, a list of emoticons with their assigned polarities and a dictionary of common phrases of the ALGD) were proposed and tested for polarity computation.

Rahab et al. (2017) proposed an approach to annotate Arabic comments extracted from Algerian Newspapers websites as positive or negative classes. For this work, they created an Arabic corpus named SIAAC (Sentiment polarity Identification on Arabic Algerian newspaper Comments). They tested two well-known supervised learning classifiers which are Support Vector Machines (SVM) and Naive Bayes (NB). For experiments, they used different parameters and various measures in order to compare and evaluate results (recall, precision and F-measure). In terms of precision, the best results were obtained using SVM and NB. It was proved that the use of bi-gramme increases the precision for the two models. Furthermore, when compared to OCA (Opinion Corpus for Arabic (Rushdi-Saleh et al., 2011)) SIAAC showed competitive results.

Guellil and Azouaou (2017) proposed an automatic parser for the ALGD which they called "ASDA" (Syntactic Analyzer of the Algerian Dialect), which labels terms in a given corpus. Their work presents a table which contains for each term its stem and different prefixes and suffixes. The goal behind such work is to help determine the different grammatical parts of a given text, in order to perform an automatic translation of the ALGD.

(Guellil et al., 2018), proposed a simple polarity calculation method for corpus annotation. It is a lexicon-based approach where the lexicon is automatically created using the English lexicon "SOCAL" (Taboada et al., 2011). Words were translated into Arabic although their polarity remained the same. The generated lexicon is then used to annotate the corpus.

It is clear from studying related works, publicly available resources for sentiment analysis in ALGD are rare. Those which are available such as (Mataoui et al., 2016), gives only the polarity of comments collected, without any information on the emotion expressed or the user expressing an opinion. It is the same as the one proposed by (Guellil et al., 2018). Therefore, we propose the first and the largest Algerian corpus annotated at both sentiment and emotion levels as well as extra-linguistic information level (age, gender, etc.).

3. Algerian Dialect Specificities and Challenges

Algerian Arabic or Algerian dialect is considered less normalised and standardised compared to MSA. It has a vocabulary inspired from Arabic, but the original words have been altered phonologically and morphologically, (Meftouh et al., 2012). Algerians express themselves in several languages, Arabic, French, English, as well as, Tamazight the original language of the first inhabitants of

[4]https://competitions.codalab.org/competitions/17751

the region. Tamazight is also divided according to regions, for example Kabyl, Chaoui, Mzabi and Tergui. More than 99% of Algerians have Tamazight and ALGD as their native language. About 73% of the country's population speak ALGD while 27% speak Tamazight[5]. The ALGD is a mixture of Turkish, Italian, Spanish, English, French, although mainly Arabic. Other new languages are also used due to culture fans for instance, Japanese, Korean and others.
It is practically the same for Tunisians and Moroccans however, Egyptians do not use as much French.
The following properties are not only specific to the Algerian dialect.

- *Code-switching:* North Africans alternate between two or more languages, or language varieties, in the context of a single conversation. This is illustrated in the following example: "C'est bon يَعتِك الصَحَا". The user has used an Arabic expression "يَعتِك الصَحَا" and a French expression "C'est bon" which means "It taste good thank you". However, the Algerian dialect is also formed by transformed words from the languages which inspired Algerians through the ages. Take the word "وذن" which is inspired from the Arabic word "أُذن" meaning "ear", where the first letter was changed. This phenomenon is known as "Intra-word switching " in linguistics, (Sankoff and Poplack, 1981), where a switch could occur in one or more places in the same word.
- *Encoding a language in letters of another language*: either Arabic expressions encoded in Roman letters known as "arabizi", or the opposite which is called "romanisation". As an example of arabizi we have "ya3tik lsaha", written in Arabic as "يَعتِك الصَحَا" meaning "thank you", and "بَاي باي" written in Arabic, which refers to the English expression "bye bye".
- *The combination of the two*: code-switching and encoding a language in letters of another one. "sba7 l5ir ça va?", an expression of a mixture of Arabic expression "sba7 l5ir : صبَاحّ الخِير" meaning "good morning" and French "ça va?" meaning "how are you?".
- *The use of numbers instead of letters or words*: this phenomenon has been observed with the proliferation of mobile phones and the social web, where users started to use more and more abbreviations. Since numbers resemble some letters and some syllables, they were used to replace those letters and syllables. Table 3. gives examples of the meaning of each number with its use.
- *Derivatives of the Algerian dialect*: It is also a fact that North Africans speak a variety of dialects in each region. In Algeria, each area is characterised by its own spoken variation of dialect. The people from Eastern and Western areas speak with totally different accents. For example the word "woman", in the East she is called "مرَا" pronounced "m'ra" in the west "ثِرَا" pronounced "sheera".

[5]https://www.worldatlas.com/articles/what-languages-are-spoken-in-algeria.html

where; (ar) : Arabic; (fr) : French

Number	It replaces	Eg : full word == meaning
3	ع	3neb : عنِب (ar) == raisins
5	خ	5ali:خَالِ (ar) == uncle
6	ط	6abla == table
7	ح	7ot: حُوت (ar) == fish
9	ق	fou9 : فَوق (ar) == over

Table 1: Which Number Replaces Which Letter?

- *Social media chats language*: social web users, especially the young, use many emoticons and emojis. Besides abbreviations (already mentioned in earlier paragraph), social media has its own language.

  Since emoticons help express emotion in a single character, its use has widely spread. "Hashtags", are used to find, follow, and contribute to a conversation. "Sharing/retweeting" a post is a way of showing support, participating or even trivialising the post.

  Another characteristic of social chatting is the use of capital letters. Internet code for Yelling and Shouting. In most cases this is considered rude. In other instances, typing everything in capitals conveys the importance of the text. There are other methods to emphasis a word or a text such as the use of *asterisks* and s p a c i n g words out or even letters' repetition to emphasise non-verbal signals (joy, anger, etc.). Letter repetition is used to overstate comments. For example: yaaaaay, stoooop[6].
- *Idioms and expressions*, which are mostly used for sarcasm, or to suggest something indirectly or covertly. For instance, "جمَال" is a way of calling someone boring, where the expression is a common name.

Above all, there is the possible existence of more than one language in the same sentence. With many possible writing styles, possible writing errors and new words, frequently appearing, makes the Algerian dialect very difficult to understand and very complex to process automatically.
These linguistic diversities call for special attention, which is why the spoken and written dialects are very rich and varied languages[7].

4. Contribution

Tools and resources are essential if progress to be made in this field of research. To ensure the credibility of resources,

[6]https://newrepublic.com/article/150506/universal-basic-income-future-of-pointless-work

[7]All words quoted from the Algerian dialect were given by the authors, who are regular users of the dialect and social media

using crowd-sourcing was considered. Hence, an open platform was created for manual annotation which we called "TWIFIL". The three main contributions of our work are:

- A crowd-sourcing annotation platform.
- Multi-grained annotations were done at both word and tweet level.
- Multi-level annotations including sentiment, emotion and extra-linguistic information (age, gender and topic).

Fig 1 presents a schema of the work detailed in this paper. As we can see the posts collected through the Twitter api are annotated where the annotators provide the annotation at both word and tweet level. Which helps create a lexicon and a corpus. These resources are then exploited to perform polarity classification.

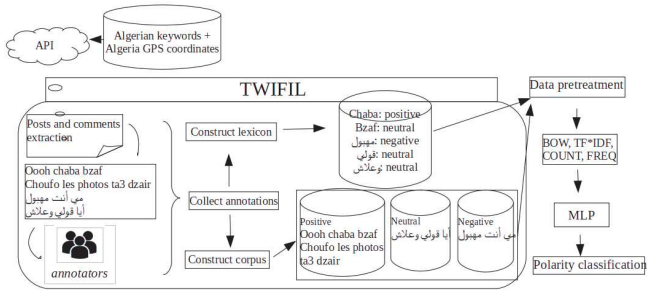

Figure 1: A general architecture of our work for OEA of the DALG

4.1. TWIFIL

TWIFIL (TWIter representing the social media and FIL of profile, meaning giving a profile to the published data, age of the author, gender, etc.) is a public platform accessible to everyone through the web[8] or mobile[9]. It was created to facilitate the generation of Algerian dialect's resources (corpus, dictionary, lexicon) but also to help researchers annotate their own data.
The annotators were given guidelines on how to annotate each text. Along with, description of each category (polarity, emotion, etc.) as well as examples of already annotated texts of the same category.
To respect users' control and privacy, texts of the tweets were the only data displayed. The administrators or the corpus holders can validate or ignore an annotation based on its consistency with regards to OEA, this was implemented to help recheck the annotations relevance.
The annotation guidelines are as follows:

- the sentiment polarity of the shared text labeled between [-10 ; +10];
- the opinion class (positive, negative, neutral);
- the emotion felt by the reader of the text labeled as (joy, anger, disgust, fear, sadness, surprise, trust, anticipation, love neutral); we followed the Plutchik eight emotion set (Plutchik, 1984) to which we added love and the neutral class to account for factual tweets;
- the topic of the text (politics, sports, diverse, etc.);
- the age of the author, labeled using age classes ([12-20], [21-30], [31-40], [41-50], [51-60], [61 and older]);
- the gender of the author (male, female, other).

The dialect lexicon is practically the same without age nor gender. Annotators provide their impressions regarding the polarity and the emotion of a word or an idiom from the ALGD.
New words can be added to the dictionary (which must be validated by an admin), different spellings added of the words and different related words. They can also add idioms with their description to facilitate the comprehension and use of the idiom.
The platform allows users to also upload their own corpora to be annotated.

4.2. The Generated Annotated Corpus

The data displayed on the TWIFIL platform are tweets collected through the Twitter API using both standard and stream. TWIFIL has more than 140k collected tweets using geo-tagging and keywords, the set of keywords contained names of known figures from politics to arts and sports, the name of some places and local events, etc. At the end we collected tweets posted between 2015 and 2019 obtained from different random geo-locations in Algeria. With the help of 26 annotators it was possible to generate a corpus and lexicon, which were validated by the admins of the platform. Considering tweets which were at least annotated by three different annotators, the labels of the corpus were assigned according to a majority vote, where a label has been used more than once otherwise the tweet will not be selected to be part of the corpus (examples can be found in Table 2).

Data statistics
As mentioned, a corpus was built of 9,000 annotated and validated tweets for sentiments. Indeed, the corpus has 4,350 positive tweets, 2,615 negative tweets and 2,191 neutral tweets. The table 3 gives the details for the tweets annotated for emotion analysis. For the age and topic we collected about 300 annotated tweets and for gender we have more than 700 (413 male, 255 female and 36 others) annotated tweets.

4.3. The Generated Annotated Lexicon

Our approach constructs a lexicon containing words in both Arabic and Latin letters with their polarity/emotion/different spellings, by using words from the lexicon proposed by (Mataoui et al., 2016) which contains 5,027 word, without considering their polarity, since we used a different scale. The lexicon was enriched by the TWIFIL users and now counts about 9,000 terms and expressions of the ALGD (examples can be found in Table 4). We followed the same approach we used during the generation of our corpus, where the labels of the words were chosen following the dominant vote.

[8] https://twifil.com
[9] shorturl.at/jntMY

Post	Polarity	Polarity class	Emotion	Age	Gender
Wa3lash tdirolna hakda khlona trankil (Why are you doing this, leave us alone)	-7	Negative	Anger	26	Male
Ch7al rahi lsa3a (what time is it)	0	Neutral	Neutral	30	male
Piii khtito kounti hayla (sister you were awesome)	5	Positive	Joy	28	female

Table 2: An excerpt of the generated corpus via TWIFIL

Joy	Anger	Disgust	Fear	Sadness	Surprise	Trust	Love	Anticipation	Neutral	Total
1,170	298	227	60	366	175	282	239	12	2,224	**5,054**

Table 3: Emotion characteristics of the corpus

5. Experiments and results

The experiments undertaken exploited the sentiment corpus and are as follows:

First, not only we implement and test SVM with different data representations (binary, frequency , etc) but we also tested the SVC (Support Vector Classification)(Chang and Lin, 2011), an adaptation of SVM for classification problems.

Second, we build a Multi-Layer Perceptron (MLP) sentiment classifier based on different neural architectures and different data representations.

Third, we explore the lexicon based methods to compare results. The lexicon-based method consists of adding two columns to the bag of words (BOW) vector. The first represents the number of negative words in the tweet, calculated using the proposed lexicon. The second represents the number of positive words which exist in the tweet.

Finally, we evaluate if deep learning models have good or higher performance for Algerian OEA than other state-of-the-art approaches.

Deep learning (DL) is a recent sub-field of machine learning and an efficient outcome of artificial neural network. In the last years, many researchers have studied DL for OEA. Since we also aim to improve the OEA of ALGD by improving the performance outcomes based on the combination of both the tested DL models and various pre-processing techniques. For this, two DL models are used, namely CNN and LSTM. We implemented the classical architecture of CNN and LSTM introduced in (Zhou et al., 2015). We have also used word embedding (WE) as part of our deep learning models. Using the Keras python library, precisely the Embedding layer [10]. It requires that the input data is digitally encoded, therefore, we used words' frequency. The Embedding layer is initialized with random weights and will learn an embedding for all of the words in the training dataset. And since recently researchers started exploring the contextual embeddings we tested the BERT, or Bidirectional Encoder Representations from Transformers. BERT, a language model introduced by Google and it has recently been added to Tensorflow hub, which simplifies integration in Keras models. We tested the BERT-Base, Multilingual Uncased.

Therefore, we separately test each of those algorithms namely SVMs, MLP classifiers, convolutional neural networks (CNN) and long short-term memory (LSTM) in ALGD.

[10]https://keras.io/layers/embeddings/#embedding

5.1. Data Pre-treatment and Methodology

Worldwide, expressed opinions and comments constitute a valuable information mine. However, the majority of the text produced by the social websites has an unstructured or noisy nature. This is due to the lack of standardisation, spelling mistakes, missing punctuation, non-standard words, repetitions and more. Indeed, such text needs a special treatment.

The purpose of this stage is to prepare the data for the following step, which is the classification of tweets such as "Oooh chaba bzaf" which translates to "ohh it is very beautiful" should be recognised as positive and the sentence "وعلَش تدِير هَكَ؟" meaning "why do you do this?" should be classified as negative. To correctly classify these sentences and others, we need to perform a set of treatments: 1) Text treatment. 2) Transformation of the texts to a machine-readable format (binary/digital).

The steps undertaken are detailed in the following;

- Filtering: replacement of URL links (e.g.http://example.com) by the term "link", Twitter user names (e.g. @pseudo - with symbol @ indicating a user name) by the term "person".
- Cleaning: removal of all punctuation marks as well as the exaggerations such as: "heyyy" replaced by "hey" and the consecutive white spaces were also removed.
- Tokenization: to segment the text by splitting it by spaces and form our BOW.
- Removing stop-words: to remove articles ("و", "نتُومَا", etc) from the BOW.

Fig 2 shows the achievement of the classifier during our experiments with and without some pre-processing treatments, where progress can be practically seen with each treatment applied separately, but also when applied to them all. The results demonstrated that pre-processing strategies on the reviews increases the performance of the classifiers. The data collected is used to extract the characteristics which will be used to train the classifier. The existence

Word/expression and different spellings	Polarity	Polarity class	Emotion
Hayla == great/هَايلَا, هَايلَى, هَايلَة	5	Positive	joy
T3ayi == boring/تعَيِ, تعَيِي, t3ay	-5	Negative	disgust
Ch7al == how much/شحَال	0	Neutral	Neutral

Table 4: An excerpt of the generated lexicon via TWIFIL

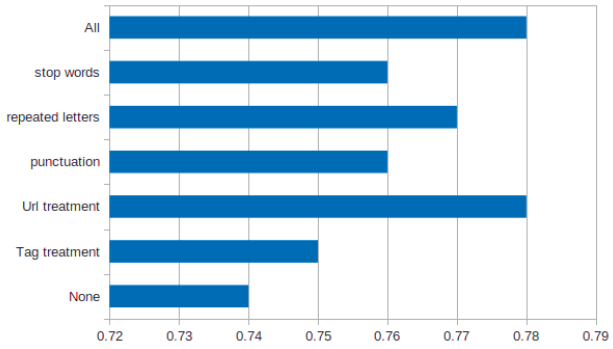

Figure 2: The evolution of the classifier's performance after each pre-processing step

of a word was used as a binary characteristic and also considered as the baseline. Tests were performed on different information representation methods, proposed in the literature, of the information retrieval field, such as the frequency of occurrence of a keyword considered as a more appropriate characteristic. During our research for approaches using this type of formatting, it was found that (Pak and Paroubek, 2010), rejected the idea and we quote "the overall sentiment may not necessarily be indicated through the repeated use of keywords". Their work was based only on a binary representation. However, others have recently used such representations (ElSahar and El-Beltagy, 2015) trained their classifiers using TF*IDF and word count. Their tests concluded that TF*IDF was the least performing method with a 3-class classification problem. However, word count gave the best accuracy, reaching 60%. In addition, (Das and Chakraborty, 2018), compared the use of TF*IDF and word existence representations, their experiments illustrated that TF*IDF is the best suited formatting for the problem. TF-IDF was used as an alternative to the binary model. However, for sentiment analysis, the binary model has been widely used by several researchers; hence, we chose to test different data representations used in the literature, namely binary, count, frequency and TF*IDF.

The result of the previous step is a vector of words, which in this step is transformed into a digital vector by: firstly, using the same dimension of the vector for all texts. Secondly: replacing the words by one of the following configurations: 1) binary 0 or 1 to represent the presence of a term. 2) Count: a simple count of the words in the text. 3) Frequency: the frequency (freq) of each word as a ratio of words within each text. 4) TF*IDF: term frequency-inverse document frequency, a statistic that reflects the importance of a word in a document, in our case the corpus.

5.2. Results and Discussion

This section presents the different results obtained, with different trained models, as well as the tests performed to choose the length of the BOW.

The corpus had about 26,000 distinct terms among which tests revealed that there are 3,000 terms, which are the most relevant terms. Such size of the vector of a tweet is what yielded the best performances in terms of accuracy (Acc).

During the experiments, we divided our corpus to three sets (training, validation and test) where 10% of the corpus was considered as the test set and 20% for the validation set and the remaining 70% constitutes our training set.

D-R	Binary	Count	**Freq**	Tf*idf	lex
SVC	67.1%	65.7%	**69.5%**	61.9%	**71%**
MLP	70.9%	68.4%	**73.4%**	68.6%	**75.3%**
CNN	68%	71%	**76%**	75%	**76%**
LSTM	71%	73%	**74%**	74%	**75%**

Table 5: The Best Data Representation (D-R) for models created for sentiment analysis

During the experiments we wanted to compare between the different SVM algorithms implemented. Experiments showed that SVC gave the best results reaching an accuracy of **69.53%** while SVM reached 63.28%.

Table 5 shows the results of the different tests performed using different data representations (D-R). The first row gives the SVC results and the second the MLP results in term of accuracy. The last column gives the results of the lexicon-based (lex) method using word frequency vectors concatenated to words' polarity count vectors. As shown, exploiting the lexicon based to create a hybrid method with machine learning yielded promising results. The same behavior was noticed whith DL models, CNN and LSTM.

Table 6 shows the top-ranked MLP architectures of the different tests performed, organised by batch size (the amount of data per training cycle), where we varied the number of epochs from 2 to 8. Looking at the number of neurons per inner layer of the network, we started with 20 neurons and reached 200. It is evident that a batch size of 200 tweets gave the best results during 2 epochs and using 180 neurons per layer.

The building of the MLP classifier was completed using a binary BOW, and we then moved on to improving its results by testing other data representations.

Table 5 illustrates that the use of other digital values such as TF*IDF or frequency can improve the accuracy of the classifier. The experiments carried out showed that the use of frequency for data encoding is the best representation for

Batch size	<50	50-200	200-400	400-600	600-800	800-1000	1000-1500
Best size	2	**200**	300	400	700	800	1400
Epoch	3/8	**2/8**	4/8	4/8	4/8	4/8	3/8
Neurons	40	**180**	100	180	180	140	100
Accuracy	67.67%	**70.86%**	69.38%	69.89%	69.23%	69.08%	68.60%

Table 6: Top-ranked MLP Architectures

the data. frequency gave the best results for MLP where an accuracy of 73% was achieved. The same applied to SVM where the best accuracy was 69.53%.

Furthermore, the exploitation of our lexicon to create a hybridisation between machine learning methods and lexicon-based methods boosted the results even further. They showed that SVM gained about 3% in accuracy, the same as MLP and LSTM, highlighted in Fig 3. CNN on the other hand, gained about 10% in term of accuracy.

To test the WE we conducted a serie of tests to choose the length of the the word vector and the results showed that a 300 length is the best for our dataset.

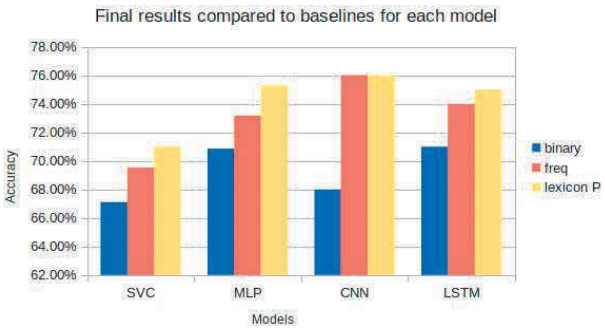

Figure 3: Final results compared to baselines (binary) for each algorithm

Through the experiment, SVM showed less performance than MLP. In fact, the best performance outputs achieved by SVM are 71% as accuracy and 75.3% for MLP.

By comparing DL and single models, the experiments show that DL enhances the efficiency of classification in terms of accuracy. In fact, the CNN and LSTM algorithms performed well and outperformed the single model (SVM and MLP). The CNN and LSTM algorithms ensure the highest accuracy with 76% and 75%, respectively.

Therefore, we can confirm that CNNs have dramatically improved the sentiment classification. One of the main differentiating factors between CNNs and traditional ML approaches is the ability of CNNs to learn to represent complex characteristics.

In table 7 we report the tests conducted on DL models. And we give accuracy results for the positive (pos), negative (neg) and neutral classes. In addition to the overall accuracy and F-measure (F1).

The best results in term of accuracy are presented in bold and were obtained with the CNN model. However, LSTM gave competitive results. On the other hand, BERT gave the worst results in term of accuracy mainly due to the out of vocabulary words. However, it gave competitive results

Model	pos	neg	neutral	Acc	F1
CNN	76%	71%	81%	**76%**	**76%**
CNN + WE	66%	46%	76%	66%	63%
LSTM	75%	68%	79%	74%	71%
LSTM + WE	77%	69%	79%	**75%**	73%
BERT	64%	58%	**82%**	68%	62%

Table 7: Deep lerning results for sentiment analysis

for the neutral class. We believe that this is due to the MSA texts present in our corpus which in general give factual information.

Considering these results, we conclude that DL models are recommended for the classification of Algerian sentiments, as they ensure high accuracy and performance compared to other methods. However, this solution has a negative effect, as it consumes more time during the training phase.

Our conclusion from these experiments confirms the conclusions obtained in other studies for Arabic and other Arabic dialects, which confirm that DL substantially improves the performance of sentiment classification (MSA (Alayba et al., 2018), Tunisian (Mulki et al., 2019), Moroccan (Oussous et al., 2019), Egyptian (Alayba et al., 2018) and Levantine (Elnagar et al., 2018)).

5.3. Error analysis

We extracted all the wrongly classified texts. After studying these texts we chose the most representative ones that are illustrated in table 8. The first example represents the examples that are positive but contains some ambiguous words like ”hungry” which represents texts that share similar vocabulary but are classified differently. The second example classified as positive while been annotated as negative. This suggests a lack of context since we don’t have enough text to know for sure.

If we look at some examples that were predicted wrongfully we understand that the most recurrent errors occur when a text contain both positive and negative words. In addition to misspellings and grammatical errors. There are also some examples that do not carry a sentiment like the third example, but were giving a positive or a negative class.

The significant phrases or words present in the texts of positive class may fall under the negative class in the training set or vice versa, which may lead to misclassification. In addition, the out-of-vocabulary problem, many words have been skipped which can also be another reason.

6. Conclusion

The Arabic language is characterised by a wide number of varieties in dialects. With the emergence of the social web,

	Text	Prediction	Actual class
1	أَسوَا احوَالنَا رُبمَا! لَكنَا وَاعُون بِعِيق نَهضَتَنَا وَلَن تضِيع جُهُود النَهضُوين سدي فَقَد أَقلَعت Our worst maybe!!But we are aware and will not hinder our renaissance and the efforts of the others will not be lost in vain	Positive	Negative
2	حنَا نفرحُو و المَسوءولِين حَاكمِين كرُوشهم we rejoice and officials governing await	Positive	Negative
3	اَلسلَام خوتِي شحَال رَاهَا دير سنديرو Hello how much is the Sadero	Negative	positive

Table 8: Error analysis of wrongly classified texts

it enables users to express their opinions using these dialects.

Algerian Dialect differs from MSA on all levels of linguistic representation, from phonology and morphology to lexicon and syntax.

Opinion and emotion analysis of the ALGD is challenging due to the rich morphology of the language. Extracting the enormous volume of comments and reviews presented on the social web requires taking into account the peculiarities of the Algerian Dialect and it's characteristics (Arabizi, code-switching, etc). Publicly available resources for OEA of the DALG are scarce.

In this paper we presented an open platform for public annotation which we called "TWIFIL". It helped create a quite large annotated corpus as well as a dialectal lexicon. These tools can be exploited for opinion and emotion analysis at a relatively low cost. This resource is now available to the community. It will provide a useful benchmark for those developing opinion and emotion analysis tools for the Algerian dialect.

As a final step, we applied various machine learning models to classify the ALGD tweets as either positive, negative or neutral. Then, we measured their accuracy and efficiency. We also analysed and evaluated the performance of the selected algorithms when applied to ALGD using different pre-processing techniques such as normalisation, stop words and URLs.

To enhance the results of the models we trained them with different data representations where term frequency proved to be more efficient than binary and TF*IDF.

To further boost the results, we used a hybridisation of machine learning models and lexicon-based methods, which surpassed the baseline results of all models. We also tested the contextual embedding using the BERT model which did not surpass our baseline.

In fact, the experimental results prove that deep learning models have a better performance for OEA of the ALGD than classical approaches (support vector machines and multi-layer perceptron).

In the future, we plan to continue with this research and address the remaining challenges, towards developing additional resources and tools for opinion and emotion analysis of Maghrebian multilingual dialects and use the obtained data to build a multilingual sentiment classifier. As well as implementing and testing other machine learning algorithms. We also plan to complete the development of the platform to allow users to add their own classes and allow the platform to offer part of speech annotations. But mainly enlarge the corpus and lexicon.

7. Bibliographical References

Abdul-Mageed, M. and Diab, M. T. (2012). Awatif: A multi-genre corpus for modern standard arabic subjectivity and sentiment analysis. In *LREC*, pages 3907–3914. Citeseer.

Abo, M. E. M., Ahmed, N., and Balakrishnan, V. (2018). Arabic sentiment analysis: An overview of the ml algorithms. In *Data Science Research Symposium 2018*, page 63.

Al-Moslmi, T., Albared, M., Al-Shabi, A., Omar, N., and Abdullah, S. (2018). Arabic senti-lexicon: Constructing publicly available language resources for arabic sentiment analysis. *Journal of Information Science*, 44(3):345–362.

Al-Smadi, M., Qawasmeh, O., Al-Ayyoub, M., Jararweh, Y., and Gupta, B. (2018). Deep recurrent neural network vs. support vector machine for aspect-based sentiment analysis of arabic hotels? reviews. *Journal of computational science*, 27:386–393.

Alayba, A. M., Palade, V., England, M., and Iqbal, R. (2018). A combined cnn and lstm model for arabic sentiment analysis. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 179–191. Springer.

Alnawas, A. and Arici, N. (2019). Sentiment analysis of iraqi arabic dialect on facebook based on distributed representations of documents. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(3):20.

Atoum, J. O. and Nouman, M. (2019). Sentiment analysis of arabic jordanian dialect tweets. *Int. J. Adv. Comput. Sci. Appl.*, 10(2):256–262.

Badaro, G., Jundi, H., Hajj, H., El-Hajj, W., and Habash, N. (2018). Arsel: A large scale arabic sentiment and emotion lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Paris, France. European Language Resources Association (ELRA)*.

Baly, R., Hajj, H., Habash, N., Shaban, K. B., and El-Hajj, W. (2017). A sentiment treebank and morphologically enriched recursive deep models for effective sentiment analysis in arabic. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 16(4):23.

Baly, R., Khaddaj, A., Hajj, H., El-Hajj, W., and Shaban, K. B. (2019). Arsentd-lev: A multi-topic corpus for target-based sentiment analysis in arabic levantine tweets. *arXiv preprint arXiv:1906.01830*.

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: A library for

support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27.

Das, B. and Chakraborty, S. (2018). An improved text sentiment classification model using tf-idf and next word negation. *arXiv preprint arXiv:1806.06407*.

Duwairi, R. M. (2015). Sentiment analysis for dialectical arabic. In *2015 6th International Conference on Information and Communication Systems (ICICS)*, pages 166–170. IEEE.

Elnagar, A., Lulu, L., and Einea, O. (2018). An annotated huge dataset for standard and colloquial arabic reviews for subjective sentiment analysis. *Procedia computer science*, 142:182–189.

ElSahar, H. and El-Beltagy, S. R. (2015). Building large arabic multi-domain resources for sentiment analysis. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 23–34. Springer.

Guellil, I. and Azouaou, F. (2017). Asda: Analyseur syntaxique du dialecte alg {\'e} rien dans un but d'analyse s {\'e} mantique. *arXiv preprint arXiv:1707.08998*.

Guellil, I., Adeel, A., Azouaou, F., and Hussain, A. (2018). Sentialg: Automated corpus annotation for algerian sentiment analysis. In *International Conference on Brain Inspired Cognitive Systems*, pages 557–567. Springer.

Habash, N., Diab, M. T., and Rambow, O. (2012). Conventional orthography for dialectal arabic. In *LREC*, pages 711–718.

Harrat, S., Meftouh, K., Abbas, M., and Smaili, K. (2014). Building resources for algerian arabic dialects.

Harrat, S., Meftouh, K., and Smaïli, K. (2017). Maghrebi arabic dialect processing: an overview. In *ICNLSSP 2017-International Conference on Natural Language, Signal and Speech Processing*.

Jarrar, M., Habash, N., Akra, D. F., and Zalmout, N. (2014). Building a corpus for palestinian arabic: a preliminary study.

Mataoui, M., Zelmati, O., and Boumechache, M. (2016). A proposed lexicon-based sentiment analysis approach for the vernacular algerian arabic. *Research in Computing Science*, 110:55–70.

Medhaffar, S., Bougares, F., Estève, Y., and Hadrich-Belguith, L. (2017). Sentiment analysis of tunisian dialects: Linguistic ressources and experiments. In *Proceedings of the third Arabic natural language processing workshop*, pages 55–61.

Meftouh, K., Bouchemal, N., and Smaïli, K. (2012). A study of a non-resourced language: an algerian dialect. In *Spoken Language Technologies for Under-Resourced Languages*.

Mohammad, S., Bravo-Marquez, F., Salameh, M., and Kiritchenko, S. (2018). Semeval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17.

Mulki, H., Haddad, H., Gridach, M., and Babaoğlu, I. (2019). Syntax-ignorant n-gram embeddings for sentiment analysis of arabic dialects. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 30–39.

Nabil, M., Aly, M., and Atiya, A. (2015). Astd: Arabic sentiment tweets dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2515–2519.

Oussous, A., Benjelloun, F.-Z., Lahcen, A. A., and Belfkih, S. (2019). Asa: A framework for arabic sentiment analysis. *Journal of Information Science*, page 0165551519849516.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326.

Plutchik, R. (1984). Emotions: A general psychoevolutionary theory. *Approaches to emotion*, 1984:197–219.

Qwaider, C., Chatzikyriakidis, S., and Dobnik, S. (2019). Can modern standard arabic approaches be used for arabic dialects? sentiment analysis as a case study. In *Proceedings of the 3rd Workshop on Arabic Corpus Linguistics*, pages 40–50.

Rahab, H., Zitouni, A., and Djoudi, M. (2017). Siaac: Sentiment polarity identification on arabic algerian newspaper comments. In *Proceedings of the Computational Methods in Systems and Software*, pages 139–149. Springer.

Rushdi-Saleh, M., Martín-Valdivia, M. T., Ureña-López, L. A., and Perea-Ortega, J. M. (2011). Oca: Opinion corpus for arabic. *Journal of the American Society for Information Science and Technology*, 62(10):2045–2054.

Saadane, H. and Habash, N. (2015). A conventional orthography for algerian arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 69–79.

Salem, F. (Feb 5, 2017). Social media and the internet of things towards data-driven policymaking in the arab world: Potential, limits and concerns. *The Arab Social Media Report, Dubai: MBR School of Government, Vol. 7, 2017. Available at SSRN: https://ssrn.com/abstract=2911832*.

Sankoff, D. and Poplack, S. (1981). A formal grammar for code-switching. *Research on Language & Social Interaction*, 14(1):3–45.

Shoukry, A. and Rafea, A. (2012). Preprocessing egyptian dialect tweets for sentiment mining. In *The Fourth Workshop on Computational Approaches to Arabic Script-based Languages*, page 47.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Zhou, C., Sun, C., Liu, Z., and Lau, F. (2015). A c-lstm neural network for text classification. *arXiv preprint arXiv:1511.08630*.

Zribi, I., Boujelbane, R., Masmoudi, A., Ellouze, M., Belguith, L. H., and Habash, N. (2014). A conventional orthography for tunisian arabic. In *LREC*, pages 2355–2361.