# A Proposed Lexicon-Based Sentiment Analysis Approach for the Vernacular Algerian Arabic

M'hamed Mataoui,[1] Omar Zelmati,[2] Madiha Boumechache[2]

[1] IS&DB laboratory, Ecole Militaire Polytechnique, Algiers, Algeria
[2] Ecole Militaire Polytechnique, Algiers, Algeria
mataoui.mhamed@gmail.com, omar.zelmati@gmail.com,
madiha.boumechache@gmail.com

**Abstract.** Nowadays, sentiment analysis research is widely applied in a variety of applications such as marketing and politics. Several studies on the Arabic sentiment analysis have been carried out in recent years. These studies mainly focus on Modern Standard Arabic among which few studies have investigated the case of Arab dialects, in this case, Egyptian, Jordanian, and Khaliji. In this paper, we propose a new lexicon-based sentiment analysis approach to address the specific aspects of the vernacular Algerian Arabic fully utilized in social networks. A manually annotated dataset and three Algerian Arabic lexicons have been created to explore the different phases of our approach.

**Keywords:** Arabic sentiment analysis, vernacular Algerian Arabic, Algerian dialect, Modern Standard Arabic, Social networks.

## 1 Introduction

The last years are mainly characterized by the fast proliferation of social networking services such as Facebook, Twitter and YouTube. These social networks allowed individuals and groups to express and share their opinions about different kinds of topics (products, political events, economics, restaurants, books, hotels, video clips, etc.). Billions of comments and reviews are added to the web each day, which has led to the need to mine users' opinions in order to discover useful information. Mining this enormous volume of comments and reviews is almost impossible manually. Therefore, a new thematic of Natural Language Processing (NLP), known as sentiment analysis (SA) or opinion mining (OM), emerged. The main purpose of sentiment analysis is to extract users' sentiments/opinions from created contents by using automatic mining techniques to determine their attitudes with respect to some topic, often expressed in textual form.

Nowadays, sentiment analysis is used mainly by businesses to discover the opinions of different customers as part of marketing purposes [1, 2]. It is also used in politics to predict election results or to know public opinions about different policies. SA field is considered as a classification task for deciding about an opinion as being positive, negative, or neutral.

According to [3], SA approaches are based on one of the two following classes: lexicon-based approaches [4, 5]; corpus-based approaches [6, 7].

Most of existing research on sentiment analysis focuses on English text [2, 4, 8]. In spite of its importance as one of the most used languages in the world, only a limited number of research on Arabic sentiment analysis has been carried out. The proposed Arabic sentiment analysis approaches mainly focus on Modern Standard Arabic (MSA) [1, 9], among which few studies have investigated the case of Arab dialects (colloquial Arabic), namely, Egyptian [1, 10-12], Jordanian [1, 9], and Khaliji (dialect used in the Gulf countries) [13]. To our knowledge, research on sentiment analysis for the Maghreb dialects or Maghrebi Arabic (Algerian, Moroccan and Tunisian) is almost non-existent [14].

The purpose of this work is to begin a reflection to study the sentiment analysis for the case of the Algerian dialect, very different compared to other Arabic dialects, not only in pronunciation, but rather by its different textual forms, very diverse and extremely rich.

This paper is organized as follows: In Section 2, related work is presented. Section 3 presents the peculiarities of vernacular Algerian Arabic. Section 4 describes our sentiment analysis approach and presents our different datasets and experimental results obtained. Finally, in Section 5, we conclude with some prospects.

## 2 Related Works

In this section, we will present research related to Arabic sentiment Analysis field with focus on dialectal Arabic study cases.

Arabic language is characterized by a wide number of dialects varieties. Besides Modern Standard Arabic used as a formal language, different Arabic dialects are used for nearly all everyday speaking situations. By the emergence of social media and the various electronic networks, enabling Arab users to express their opinions using different Arabic dialects, researchers have raised the need to consider this amount of generated content especially by the study of the peculiarities related to written forms of these different dialects.

Research on Arabic sentiment analysis field can be classified into three categories: First, the work that are interested in building SA related resources (corpus and lexicons). Secondly, the work which propose sentiment analysis approaches for MSA and Arabic dialects (lexicon-based, corpus-based). Finally, work which propose techniques related to SA improvement (pre-processing, morphological analyzers, etc).

### 2.1 Building resources for Arabic sentiment analysis

Abdul-Mageed and Diab [15] constructed a large-scale multi-lingual lexicon based on both MSA and colloquial Arabic (Egyptian and Levantine) for sentiment analysis, called SANA. SANA lexicon is a combination of many lexicons, such as, SIFAAT, HUDA and an automatic collected corpora (with both statistical method and machine translation).

Abdulla Nawaf et al. [13] addressed the issue related to the lack of freely accessible datasets for analysis and testing in the Arabic sentiment analysis context. A relatively large dataset of Arabic comments and reviews from Yahoo!-Maktoob social network has been manually collected and annotated. The annotation phase was made by two or three experts in MSA and Jordanian dialect. Authors have used two classifiers (SVM and Naive Bayes) in their experiments. They have showed that SVM outperforms NB and achieves about 64% of accuracy level [13].

Gilbert Badaro et al. [16] addressed issues related to the build of an Arabic sentiment lexicons. They produce ArSenL, the first publicly available large scale Arabic sentiment lexicon. ArSenL is based on a combination of existing resources, like: ESWN, Arabic WordNet, and SAMA. Their experimental study shows that using English-based linking produces superior performance in comparison to using the WordNet-based approach. Authors showed also that the combination of the two resources is better than either.

Diab et al. [17] developed an electronic lexicon that can be used in different NLP tasks, sentiment analysis in our case. Their lexicon consists of three parts: MSA, dialectal Arabic and English. Authors made Tharwa publicly available which can be used mainly for the Egyptian dialect sentiment analysis.

Al-Kabi et al. [18] shows the creation of a flexible and relatively big corpus, that consists of 250 topics equally divided among five domains (economy, food-life style, religion, sport, and technology), for Arabic sentiment analysis. Their manually created corpus is characterized by its flexibility and is constituted mainly of comments and reviews expressed in both MSA and Colloquial Arabic. It contains five types of reviews (English, mixed MSA & English, French, mixed MSA & Emoticons, and mixed Egyptian & Emoticons). Authors show in their analysis that most of the users of Yahoo! Maktoob prefer to use of MSA. In addition, the created corpus contains few comments and reviews that used English, French, Emoticons, etc.

### 2.2 Sentiment analysis approaches for MSA and Arabic dialects

Itani et al. [19] conducted a comparison between the lexicon-based and corpus-based approach by using both MSA and Arabic dialects. The experimental results shows that lexicon-based approach (83.4% of accuracy) outperforms the corpus-based approach.

Ahmad et al. [20] are the first who investigated Arabic sentiments analysis by studying the case of financial news. They showed that the proposed local grammar approach, developed on an archive of English texts can be applied to both Chinese and Arabic languages.

El-Beltagy et al. [21] focused on problems, challenges and open research issues related to Arabic sentiment analysis. They proposed to build domain-based and vernacular-based Arabic sentiment lexicons and consider the computation of the semantic orientation of Arabic Egyptian tweets as a case study for which an Egyptian dialect sentiment lexicon has been created. They used two methods of polarity computation: straightforward sum and double polarity sum. Experimental results showed that the use of weighted lexicons with double polarity sum obtained good improvements.

Al-Kabi et al. [1] proposed to build a sentiment analysis tool for colloquial Arabic and MSA, called: CNSA-MSA-SAT. They collected a large number of comments and reviews to build polarity lexicons used by CNSA-MSA-SAT tool. Authors also built 18 specialized polarity lexicons for both colloquial Arabic and MSA. Considered polarities are: positive, negative, and neutral. Experimental results showed that CNSA-MSA-SAT tool obtained an accuracy rate of 90% over the test dataset.

Al-Kabi et al. [22] developed an opinion mining and analysis tool to evaluate Arabic social content, both colloquial and MSA. Comments and reviews are evaluated according to three characteristics: subjectivity (subjective or objective), Polarity (positive or negative) and weight (strong or weak). Experimental results of the developed analysis tool showed that the proposed approach obtained more accurate results when it is applied on specific domain reviews (politics, technology, products, etc.).

Abdulla Nawaf et al. [23] have opted for a lexicon-based unsupervised sentiment analysis approach with a manual creation of the lexicon. Their approach is based on two components: the lexicon and the SA tool. The test dataset was created from two corpora each prepared separately. The first consists of 2000 tweets (1000 positive and 1000 negative) written in MSA and Jordanian dialect. The second corpus, extracted from Yahoo Maktoob, collected to meet the same criteria as the Twitter corpus (2000 comments with 1000 positive and 1000 negative). Authors obtained a low accuracy level in experiments, thus, they suggest some improvements by: expanding the lexicon, including the concept of weighted positive/negative sentiment for each word of the lexicon, proposing new improved combination techniques in the overall polarity computation, carrying out experimentations on bigger and diverse datasets, etc.

Hossam S. Ibrahim et al. [12] presented a feature-based sentence level sentiment analysis approach for Arabic language. They used a lexicon consisting of Arabic phrases to improve the polarity detection of Arabic sentences. Also, many linguistic features have been used, such as, Intensifiers, Shifters and negation. The developed lexicon focuses on both MSA and Egyptian dialectal Arabic. Experimental results showed that the proposed approach obtained 95% of accuracy using SVM classifier.

### 2.3 Related techniques to improve Arabic sentiment analysis

Shoukry and Rafea [11] studied the effect of pre-processing mechanisms on the performance of an Arabic sentiment analysis. Authors used a dataset consisting of 1000 tweets expressed in Egyptian Arabic dialect extracted from Twitter. They have used two stemmers over two approaches (Machine Learning (ML) and Sentiment Orientation (SO)). By using the pre-processing module combined with their stemmer, authors obtained improvement of 4.5% (respectively between 2-7%) for ML approach in all used measures (respectively for SO approach).

In [10], Shoukry and Rafea proposed an implementation of a sentiment classification for Arabic tweets. They investigated the use of the machine learning approach for Arabic sentence level sentiment analysis by using 1000 extracted tweets. Two Classifiers have been used: Naïve Bayes and SVM. Authors mentioned some problems related to the training corpus which could affect the classification accuracy.

Al-Kabi et al. [24] conducted a comparative study between two sentiment analysis tools, SocialMention and Twendz, by using a dataset containing 4,050 Arabic and English reviews collected from Yahoo news, YouTube, Facebook, Twitter, etc. Three polarity dictionaries (Arabic, English and emoticons) have been manually constructed based on this dataset. The experimental results has shown that SocialMention is more accurate to identify the polarity of Arabic/English comments compared to Twendz.

Salloum, Wael, and Nizar Habash [25] propose morphological analyzers for dialectal Arabic (called ADAM). They extend an MSA morphological analyzer's database through a set of handwritten rules to add new entries of dialectal affixes into this database. Experimental evaluation showed that ADAM has decreased to half the rate of out-of-vocabulary compared to SAMA.

Sadat et al. [26] developed a framework for Arabic dialects classification using probabilistic models across social media datasets. They carried out a set of experiments exploiting the n-gram technique with Markov language model and Naive Bayes classifiers. These experiments showed that Naive bayes classifier based on bi-gram model was able to get very good results by identifying 18 different Arabic dialects with an accuracy rate of 98%.

Saadane Houda and Nizar Habash [27] presented a basic layout of Algerian Arabic processing. This layout can be used in most of NLP applications, such as sentiment analysis. The authors carried out a comparison with other Arabic dialects (Egyptian, Tunisian, etc.).

## 3 Algerian Dialect

Algerian Arabic or Algerian dialect (ALGD) is considered as one of the most "hard to understand" Arabic dialects varieties. It is far less normalized and standardized compared to MSA. It has a vocabulary inspired from Arabic but the original words have been altered phonologically [28]. ALGD belongs to Maghrebi Arabic (Western group) and is mainly used in daily life. It is characterized by the absence of writing resources, hence it is considered as an under resourced language [27]. ALGD differs from MSA and other Arabic dialects by having many specific features. In addition to MSA and dialectal Arabic, a rich vocabulary consisting of foreign words of French origin are an essential part of the spoken language of Algerians.

Phonology, morphology, lexicon and syntax of ALGD are very difficult to understand for the citizens of the other Arab countries.

For historical reasons, ALGD has been enriched by many languages (Turkish, Italian, Spanish and mainly by French) which resulted a complex linguistic situation.

With the advent of social networks, the ALGD is increasingly used by the Algerian Web users. According to ITU[1], 28% of Algerians are actively using Internet. Most of this activity is dominated by using social networks. Millions of comments and reviews are added every day. Mining this enormous volume of comments and reviews

[1] International Telecommunication Union

requires taking into account particular aspects of ALGD. Thus, our use of ALGD will focus mainly on the written form and its characteristics.

The first feature of ALGD is the use of words that comes from several languages (Code-Switched). Algerian vernacular Arabic is often known as a dialect code-switched with French [29]. To illustrate this feature, we can give the example of a comment excerpted from our test corpus: "top 444 ربي يوفقك" in which the user has used the words "ربي" and "يوفقك" that are of Arab origin, and the words "top" and "444" (4 is expressed to represents the word fort which means strong) that are of French origin. This comment means "top, strong, god helps you ".

The second feature is related to the use of Arabic expressions encoded in Romanized Arabic or foreign expressions (mostly French) encoded in Arabic letters. As example of the first case (Arabic encoded Romanized, known also as *arabizi*), we mention the comment: " *itar kbir flblad w ..... ysab Addine Hadiya karitha* ", equivalent to the following expression in Arabic: "إطار كبير فالبلاد ويسب الدين.... هذي كارثة", which means: *"A senior executive of the state, and he insults the religion ... it's a disaster"*. For the second case (French encoded in Arabic letters), we mention the comment from our test corpus:"جامي نيتيليزي سكايب", that represents the French expression : "je n'utiliserai jamais skype", which means:" I will never use skype ".

The third feature is the combination of the two first features, i.e. code-switched expressions containing words encoded with Romanized Arabic mixed with French words (or other foreign languages) encoded in Arabic letters. We mention this example from our test corpus:"khorda عرة الطوموبيلات ". This expression contains an Arabic word of the Algerian dialect written in Arabic letter (عرة which means worst), a modified French word written in Arabic letters (الطوموبيلات which means cars) and a Romanized Arabic word (khorda which means scrap). This expression means "the worst car, scrap".

The last feature is related to the use of words written in a very specific form, the form that most Algerians generally used for writing short messages. For instance, the word "mli7" (which represents the Algerian Arabic word "مليح") which means "Good". Also, the word "3ayane" (which represents the Algerian Arabic word "عيان") which means in certain cases "tired" or "bad" in other cases. We note here the use of Arabic numerals to present Arabic letters that are close in their writing "7" for "ح", "3" for "ع", etc. and the use of abbreviations.

All these features make the spoken and written Algerian dialect a very rich and varied language which requires special consideration of all of these properties and linguistic diversity.

## 4 The Proposed Lexicon-Based Sentiment Analysis Approach

This section exhibits our proposed lexicon-based sentiment analysis approach. Our approach attempts to address several issues related to sentiment analysis for the specific case of Algerian vernacular Arabic. These problems mainly lie in:

- All features mentioned in section 3, related to specific aspects of ALGD;
- Unavailability of Algerian vernacular Arabic sentiment lexicons;

- Unavailability of Algerian vernacular Arabic parsers.
- Unavailability of test dataset for the Algerian vernacular Arabic content.

To address the first problem, i.e. related to specific aspects of ALGD mentioned in section 3, we propose a process to handle each of these aspects. Our process is based on four modules: common phrases similarity computation module; pre-processing module; language detection & stemming module; and polarity computation module.

To address the second problem, we built three lexicons which are: keywords lexicon (L1); negation words lexicon (L2); intensification-words Lexicon (L3). Two other resources were used: a list of emoticons with their assigned polarities, and a dictionary of common phrases of the ALGD.

Concerning the third point, namely Algerian vernacular Arabic parsers, we have implemented a parser which takes its strength from our knowledge of different forms of expression used by the Algerian Web users. Our parser is based on the three following steps: tokenization, normalization and stop-words removal.

To address the last point, i.e. test dataset problem, we collected "post and comments" during a period of about a month from several pages of the Facebook social network very well-known and frequented in most cases by Algerian Web users. These dataset has been filtered and annotated by experienced users to form our test corpus.

### 4.1 Algerian Vernacular Arabic Sentiment Lexicons

As aforementioned, our sentiment analysis approach is based on three lexicons:

- Keywords lexicon (L1);
- Negation words lexicon (L2);
- Intensification words Lexicon (L3).

To build our L1 lexicon, we relied on the work of the text mining research group at Nile University2 in which they set a lexicon containing the words and expressions in Arabic and Egyptian dialect annotated with their polarities. We firstly remove all words and expressions not used in the ALGD. After that, we have included all the words (with their respective polarities) of the ALGD equivalent to Arabic and Egyptian words. Finally, we added the words of the Algerian dialect commonly used to express positive or negative opinion.

At the end of these steps, our lexicon L1 was composed of 2380 words with a negative polarity and 713 with a positive polarity.

For the other two lexicons, we used an MSA dictionary of negation and intensification words. So, we added all equivalent words used in the ALGD to these lexicons.

As aforementioned, two other resources were prepared: a list of emoticons and a dictionary of common phrases of the ALGD with their assigned polarities.

[2] http://tmrg.nileu.edu.eg/

### 4.2 Lexicon-Based Sentiment Analysis Process for ALGD

Our lexicon-based sentiment analysis process is based on four modules:

- common phrases similarity computation module;
- pre-processing module;
- language detection & stemming module; and
- Polarity computation module.

Our approach may involve other work to better define the nature of expressions before processing. As an example we can cite the work of Sadat et al. [26] allowing the automatic identification of Arabic dialects.

**Common phrases similarity computation module.** The first module, i.e. common phrases similarity computation module, allows to deal with common expressions before passing at the word level handling. This module compare the input text (comment) with the "common phrases table" by computing its score of similarity (N-gram similarity). If the score of similarity exceeds a certain threshold, the module will consider the input text as a common phrase, therefore, no need to proceed to the word by word handling.

**Pre-processing module.** The pre-processing phase is very important for all NLP tasks. Shoukry and Rafea [11] indicated that this phase has a very positive impact on the performance of sentiment analysis.

The pre-processing module is mainly based on our parser, it extracts the tokens (keywords, negation words, intensification words and emoticons) by proceeding according to the following steps: tokenization, normalization and stop-words removal.

Arabic language is known by the property to have multiple forms of a given letter, for instance, "ا، أ، إ، آ، ا" are several forms for the letter "ا" (alif). Therefore, the normalization step serves to transform every letter to its defined standard form.

For the stop-words removal step, it consists often to remove common words that are unrelated to the topic of interest, such as "and" and "the" in English language. In the information retrieval (IR) field, this step is very important in both indexing and retrieval phases. All stop words are removed by the IR system. Contrary to this, some stop words can play an important role in SA field. For instance, we can cite the case of following stop words: "مع" and "و". The first word expresses an agreement, as used in the phrase: "أنا مع هذا الإقتراح", which means "I agree with this proposal". The second word, i.e "و", is used as a linking word between two sentences. In most of the cases, this word is used to link sentences referring to two aspects of the same topic. For instance, the expressions "هذا تلفون جميل و لكنه غالي", which means "this phone is beautiful but expensive ". For this, we have defined a list of stop words a bit limited compared to that used in the field of information retrieval.

The output of this module consists of a list of tokens (encoded in Arabic, Romanized or emoticons).

**Language detection & stemming module.** The processing performed by this module is detailed figure 1. It takes as input the results of the previous module. The first step of this diagram is to detect the language of a token $T_i$. If $T_i$ belongs to Arabic, then the module will calculate its stem using a light stemming Arabic tool, for instance, khoja stemmer[3]. The other case is when $T_i$ don't belongs to Arabic, we will have two sub-cases. The first sub-case is when a word belongs to another language. In this sub-case the module will carry out the translation of the word to Arabic encoded. As example, we can mention the word "formidable" that will be translated to "رائع". In the second sub-case, a specific translation is needed here, i.e. suggestion. For instance, the word "مليح" will be suggested by the translator (Google in our experiments) to replace the word "mlih" (Romanized Algerian Arabic word), which means "good".

At the last stage of this module, we will have as result a list of stemmed tokens.

**Polarity computation module.** This section describes (through the diagram of Fig. 2) the way our module computes the sentiment orientation of each term, and therefore aggregates these terms SO to obtain the SO of the entire text. After initialization of the text sentiment orientation (TSO) to 0, the first step of the polarity computation process consists in verifying the membership of the term $T_i$ to the lexicon L1 (words with polarities, see section 4.1). The terms belonging to L1 are of three polarities (positive, neutral and negative). If the current term does not belong to L1, the module will process the next term of the text. Otherwise, a set of rules will be checked, primarily involving interaction with words belonging to the L2 and L3 lexicons.

The first rule attempts to verify if a term $T_i$ is is preceded by a term $T_k$ and succeeded by a term $T_j$, where $T_k$ belongs to L2 and $T_j$ belongs to L3. If this is the case, the value "$Weight(T_k)* SO(T_i)$" will be added to the *TSO*.

As example of this rule, we can mention the case of the phrase: "ماشي مليح بزاف", which means "it is not so good".

As example of the second rule, we can mention the subexpression "مليح بزاف", which means "So good".

The third rule processed in this module is related to the negation case. For instance, the subexpression "ماشي مليح", which means "Not good". If this rule is verified, the module will add the value "$Weight(T_k)* SO(T_i)$" to the *TSO*. If not, only the SO of the term $T_i$ will be added to *TSO*.

### 4.3 Test Corpus

In this section, we will explore the different characteristics of our test corpus. This corpus will be used in the experimental phase of the present research. This step consisting in collecting and annotating the dataset (assigning a polarity to each comment) is very expensive in terms of both time and effort.

[3] See: http://zeus.cs.pacificu.edu/shereen/research.htm

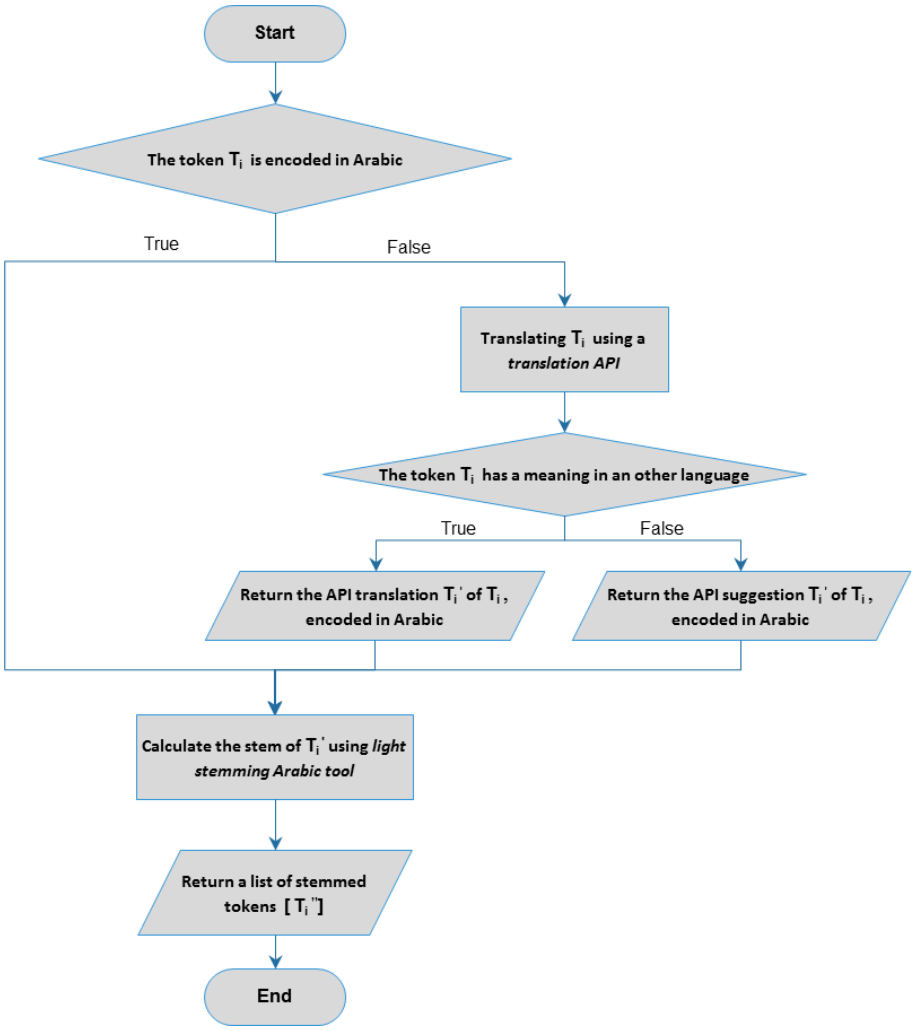


**Fig. 1.** Processing diagram of the language detection & stemming module

**Data collection.** Our data (post and comments) are exclusively extracted from Facebook. This is justified by the high use of this social network by Algerian Web users (more than 96%[4]). According to Facebook, 11M Algerians use this social network.

We have developed a module based on Facebook4J[5] for the data extraction from the Facebook social network.

[4] http://gs.statcounter.com/#social_media-DZ-monthly-201501-201601-bar
[5] http://facebook4j.org/

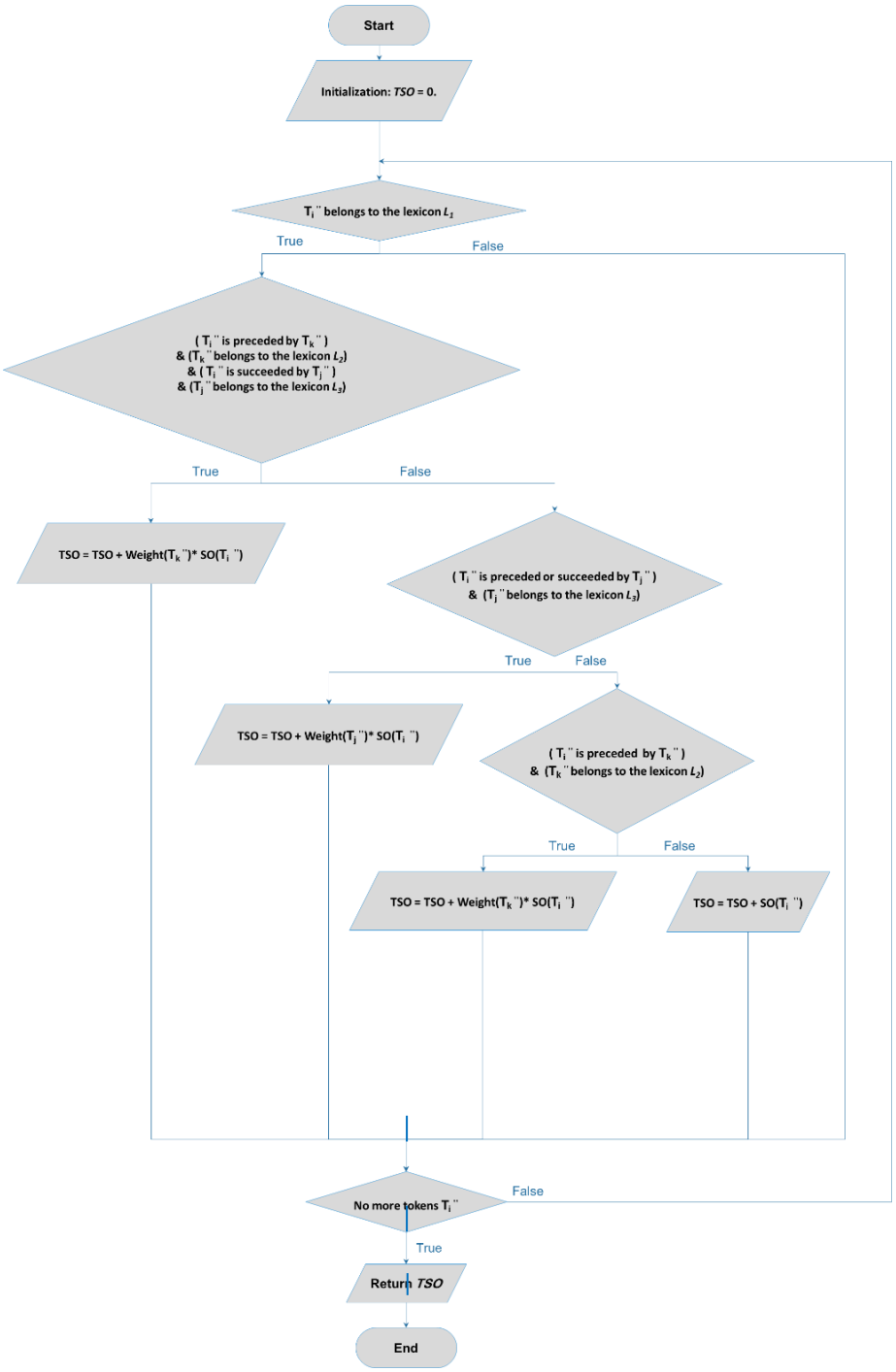


**Fig. 2.** Processing diagram of the polarity computation module

In order to target the right content, we made use of data provided by socialbakers[6]. These data show the main frequented Algerian pages. Thus, we have chosen the following pages:

- "احلام مستغانمى", a writer with more than 9,154,070 fans;
- "خديجة بن قنة", a journalist with more than 7,430,493 fans;
- "Lotfi DK", a rap songer with more than 4,409,397 fans;
- "Ooredoo", a telecom operator with more than 2,972,330 fans.

etc. These pages discuss various areas of life, i.e. economy, politics, literature and arts, etc.

The collected data has been filtered and annotated by experienced users to form our test corpus.

**Dataset Properties.** In total, we selected 206 posts comprising 7698 comments. . As aforementioned, all these comments were assessed manually by defining their polarities. The collected data discuss various areas of life from which we can mention the pages: economy ("Ooredoo", "Djezzy", "Mobilis", "Samsung Algérie"), politics ("Ali Benflis", "عبد العزيز بوتفليقة"), society ("Algérie"), literature and arts ("احلام مستغانمى", "خديجة بن قنة", "DZjoker"), sports ("الفريق الوطني الجزائري", " Maracana"), divers ("Hafid Derradji", "Journal el Bilad", "Zinou Kds", "fibladi.com", "Karim El Gang", "Echorouk online", "1.2.3 viva l'algerie", "El khabar"), etc. Table 1 shows the distribution of the collected data according to their topics.

**Table 1.** Distribution of the collected data according to their topics

| Topic | number of posts | number of comments |
|---|---|---|
| economy | 68 | 1705 |
| politics | 33 | 2422 |
| society | 32 | 1263 |
| literature and arts | 49 | 1215 |
| divers | 24 | 1093 |

As it can be seen from Table 1, the posts belonging to politics and society topics has the highest frequency of comments.

In Table 2, we present data from a perspective of the lexicon and encoding used. We note that most of comments (56%) use the ALGD encoded in Arabic or Romanized. Foreign comments are mainly words from the French.

**Table 2.** Number of comments according to thier Encoding class

| | Number of comments |
|---|---|
| MSA encoded with Arabic letters | 1503 |
| MSA encoded with Romanized letters | 36 |
| ALGD encoded with Arabic letters | 2429 |

[6] http://www.socialbakers.com/statistics/facebook/pages/total/algeria/

| | |
|---|---|
| ALGD encoded with Romanized letters | 1909 |
| Foreign languages encoded with Romanized letters | 1281 |
| Foreign languages encoded with Arabic letters | 7 |
| MSA and/or ALGD encoded with Arabic letters | 312 |
| MSA and/or ALGD and/or Foreign languages encoded with Arabic letters and/or Romanized letters | 221 |

We present in the following table some examples of comments from our test corpus according to their encoding classes (aforementioned in Table 2).

**Table 3.** Examples of comments from our test corpus

| Original comment | Translated Comment | Encoding class |
|---|---|---|
| والله صحيح, شكرا أحلام على النصيحة | Right, Thanks Ahlam for the advice | MSA encoded with Arabic letters |
| Sabah elward, kalam jamil | Good morning, nice words | MSA encoded with Romanized letters |
| راك فور خو | You are strong brother | ALGD encoded with Arabic letters |
| ma 3andehomch anti derapage hhhh | they cannot change their minds | ALGD encoded with Romanized letters |
| Facebook devrait créer le bouton "J`adore" Lotfi DK. | Facebook should create the button "I love" Lotfi DK. | Foreign languages encoded with Romanized letters |
| تووووووب | Top | Foreign languages encoded with Arabic letters |
| بالتوفيق و لعقوبة لنجاحات أخرى . | Good luck, And other successes. | MSA and/or ALGD encoded with Arabic letters |
| bonne chance rabi m3akoum | Good luck, god helps you | MSA and/or ALGD and/or Foreign languages encoded with Arabic letters and/or Romanized letters |

### 4.4 Experimental Results

In this section we present the obtained experimental results. Experiments were conducted based on our constructed ALGD dataset by using classic precision measure (accuracy).

From results of table 4, we can observe that the best configuration of our experiments is related to the use of the combination: "Arabization + Translation + khoja Stemmer".

To test the impact of the "common phrases similarity computation module", we have defined two configurations (with and without this module). According to table 5, the obtained results show that this module allowed us to improve the accuracy of our system.

**Table 4.** Impact of arabization, translation and stemming phases

| | Basic analyzer | with Arabization | with Arabization + Translation | with Arabization + Translation + Light Stemmer | with Arabization + Translation + Khoja Stemmer |
|---|---|---|---|---|---|
| Accuracy | 53.3% | 65.0 % | 71.9 % | 72.05% | 76.68 % |

**Table 5.** Results obtained by the two configurations related to the "common phrases similarity computation module"

| | Without using "common phrases similarity computation module" | By using "common phrases similarity computation module" |
|---|---|---|
| Accuracy | 76.68 % | 79.13 % |

## 5 Conclusion

We proposed in this paper a new lexicon-based approach for vernacular Algerian Arabic sentiment analysis. This approach attempts to address the specific aspects of this very particular Arabic dialect. All these aspects that were apparent before that in spoken language, but now with the advent of social networks these features exist throughout the generated content of Algerian Web users.

We mentioned in this work the main issues related to these features and proposed an approach composed of four modules: common phrases similarity computation module; pre-processing module; language detection & stemming module; and polarity computation module. Our built lexicon is composed of three parts: keywords lexicon; negation words lexicon; intensification words lexicon. These three lexicons are enriched by a dictionary of emoticons and another dictionary of common phrases.

Finally, we have built a test corpus for experimental purposes. This corpus was filtered and annotated in order to facilitate the evaluation process of our proposal.

Experimental results show that our system obtains good performance with 79.13% of accuracy.

## Acknowledgment

We thank all members of the "*Text mining research group*" of *Nile University* for giving us the opportunity to exploit their data sets.

## References


1. Al-Kabi, M., et al. *An opinion analysis tool for colloquial and standard Arabic*. in *The fourth International Conference on Information and Communication Systems (ICICS 2013)*. 2013.
2. Pang, B. and L. Lee, *Opinion mining and sentiment analysis.* Foundations and trends in information retrieval, 2008. **2**(1-2): p. 1-135.
3. He, Y. and D. Zhou, *Self-training from labeled features for sentiment analysis.* Information Processing & Management, 2011. **47**(4): p. 606-616.
4. Taboada, M., et al., *Lexicon-based methods for sentiment analysis.* Computational linguistics, 2011. **37**(2): p. 267-307.
5. Ding, X., B. Liu, and P.S. Yu. *A holistic lexicon-based approach to opinion mining*. in *Proceedings of the 2008 International Conference on Web Search and Data Mining*. 2008. ACM.
6. Kumar, A. and T.M. Sebastian, *Sentiment analysis on twitter.* IJCSI International Journal of Computer Science Issues, 2012. **9**(3): p. 372-378.
7. Klenner, M., S. Petrakis, and A. Fahrni. *Robust Compositional Polarity Classification*. in *RANLP*. 2009.
8. Pak, A. and P. Paroubek. *Twitter as a Corpus for Sentiment Analysis and Opinion Mining*. in *LREC*. 2010.
9. Duwairi, R.M., et al. *Sentiment Analysis in Arabic Tweets*. in *Information and Communication Systems (ICICS), 2014 5th International Conference on*. 2014. IEEE.
10. Shoukry, A.M., *Arabic sentence level sentiment analysis*. 2013, The American University in Cairo.
11. Shoukry, A. and A. Rafea. *Preprocessing Egyptian Dialect Tweets for Sentiment Mining*. in *The Fourth Workshop on Computational Approaches to Arabic Script-based Languages*. 2012.
12. Ibrahim, H.S., S.M. Abdou, and M. Gheith, *Sentiment Analysis For Modern Standard Arabic And Colloquial.* arXiv preprint arXiv:1505.03105, 2015.
13. Abdulla, N.A., M. Al-Ayyoub, and M.N. Al-Kabi, *An extended analytical study of arabic sentiments.* International Journal of Big Data Intelligence 1, 2014. **1**(1-2): p. 103-113.
14. Elkhlifi, A. and R. Bouchlaghem, *Opinion Extraction in Moroccan Dialect Texts.*
15. Abdul-Mageed, M. and M.T. Diab. *SANA: A Large Scale Multi-Genre, Multi-Dialect Lexicon for Arabic Subjectivity and Sentiment Analysis*. in *LREC*. 2014.
16. Badaro, G., et al., *A large scale Arabic sentiment lexicon for Arabic opinion mining.* ANLP 2014, 2014: p. 165.
17. Diab, M., et al. *Tharwa: A large scale dialectal arabic-standard arabic-english lexicon*. in *Proceedings of the Language Resources and Evaluation Conference (LREC)*. 2014.
18. Al-Kabi, M., et al., *A Prototype for a Standard Arabic Sentiment Analysis Corpus*, in *The International Arab Conference on Information Technology (ACIT'2015)*. 2015: Amman, Jordan.
19. Itani, M.M., et al. *Classifying sentiment in arabic social networks: Naïve search versus Naïve bayes*. in *Advances in Computational Tools for Engineering Applications (ACTEA), 2012 2nd International Conference on*. 2012. IEEE.

20. Ahmad, K., D. Cheng, and Y. Almas. *Multi-lingual sentiment analysis of financial news streams*. in *Proc. of the 1st Intl. Conf. on Grid in Finance*. 2006.
21. El-Beltagy, S.R. and A. Ali. *Open issues in the sentiment analysis of Arabic social media: A case study*. in *Innovations in information technology (iit), 2013 9th international conference on*. 2013. IEEE.
22. Al-Kabi, M.N., et al., *Opinion mining and analysis for arabic language.* IJACSA) International Journal of Advanced Computer Science and Applications, 2014. **5**(5): p. 181-195.
23. Abdulla, N.A., et al., *Towards improving the lexicon-based approach for arabic sentiment analysis.* International Journal of Information Technology and Web Engineering (IJITWE), 2014. **9**(3): p. 55-71.
24. Al-Kabi, M., et al. *Arabic/English sentiment analysis: an empirical study*. in *The Fourth International Conference on Information and Communication Systems (ICICS 2013)*. 2013.
25. Salloum, W. and N. Habash, *ADAM: Analyzer for Dialectal Arabic Morphology.* Journal of King Saud University-Computer and Information Sciences, 2014. **26**(4): p. 372-378.
26. Sadat, F., F. Kazemi, and A. Farzindar. *Automatic identification of Arabic dialects in social media*. in *Proceedings of the first international workshop on Social media retrieval and analysis*. 2014. ACM.
27. Saadane, H. and N. Habash. *A Conventional Orthography for Algerian Arabic*. in *ANLP Workshop 2015*. 2015.
28. Meftouh, K., N. Bouchemal, and K. Smaïli. *A study of a non-resourced language: an Algerian dialect*. in *SLTU*. 2012.
29. Cotterell, R., et al. *An algerian arabic-french code-switched corpus*. in *Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Programme*. 2014.