

# An Enhanced Approach for Arabic Sentiment Analysis

Haidy H. Mustafa<sup>1</sup>, Ammar Mohamed<sup>2</sup>, and Doaa S. Elzanfaly<sup>3</sup>

<sup>1</sup>Faculty of Computers and Information, Helwan University, Cairo, Egypt & Department of Computer Science, ISSR, Cairo University, Egypt

<sup>2</sup>Department of Computer Science, ISSR, Cairo University, Egypt

<sup>3</sup>Faculty of Computers and Information, Helwan University, Cairo, Egypt & Informatics and Computer Science, British University in Egypt, Cairo, Egypt

## ABSTRACT

*In recent years, there are massive numbers of users who share their contents over wide range of social networks. Thus, a huge volume of electronic data is available on the Internet containing the users' thoughts, attitudes, views and opinions towards certain products, events, news or any interesting topics. Therefore, sentiment analysis becomes a desirable topic in order to automate the process of extracting the user's opinions. One of the widely content sharing languages over the social network is Arabic Language. However Arabic language has several obstacles that make the sentiment analysis a challenging problem. Most users share their contents in informal Arabic. Additionally, there are lots of different Arabic dialects. Hence, Arabic sentiment analysis researches is developed slowly compared to other languages such as English. This paper proposes a new hybrid lexicon approach for Arabic sentiment analysis that combines in the same framework both unsupervised and supervised technique. In the unsupervised phase, the polarity of data is extracted by means of Look-up table stemming technique. In the supervised phase, we use the data of the true classified polarity from the unsupervised phase to generate and train a classifier for the further classification of the unclassified data. We test and evaluate the proposed approach using MIKA corpus [1]. The results show that the proposed approach gives better results.*

## KEYWORDS

*Sentiment Analysis, Look-up Table, Stemming, Arabic Social Media.*

## 1. INTRODUCTION

Web 2.0 [2] allows users to interact, collaborate and share the content across several social media web sites. Around 6 billion photos are uploaded monthly to Facebook with blogosphere doubles every five months, 72 hours of videos are uploaded every minute on YouTube, and more than 400 million daily tweets on Twitter [3]. These huge amounts of data reflect the user preferences, thoughts, opinions, ideas, and etc. Understanding and analyzing these data can be very useful for business market or governments. For example, in business market analyzing social data can help us to estimate the customers' needs (opinions) and impressions towards services or products. In government, investigating the impression of users' data help them to take a suitable decision. The process of extracting and analyzing user's opinions, attitude, and preferences is called sentiment analysis [4]. Broadly speaking, sentiment analysis aims at determining the appraisal attitude and emotions towards entities such as products, services, organizations, individuals, issues, events, topics.

Generally, social media users share content in different languages. The widely used language for content sharing is English. Thus, most of available researches in Sentiment Analysis are made in English. One of the other different languages for content sharing is Arabic language. Arabic language is one of six official languages of the United Nations [5]. Additionally, Arabic is the

official language of 21 countries, and it's the major language in several areas of the world [6]. The number of Arabic social media users is annually increasing. According to the Arab Social Media Report in [7], the number of Facebook users increased from 115 million to 156 million in the period from 2016 to February 2017. Also, the total number of active Twitter users in the Arab world reached around 11.1 million by March 2017. The total number of LinkedIn users in the Arab world reached around 16.6 million by January 2017. This rapid growth of Arabic users over different Social networks yields to continuous increase in Arabic shared opinions. This asks for an automated method to extract the users' opinions.

Broadly speaking, the Arabic language possess some characteristics. Arabic is Semitic language, and it's written from right to left. Arabic words are formed from root. In addition, the words encompass very complicated derivatives and inflections. In fact, Arabic language is classified into three types; Classical, Modern Standard, and Colloquial Arabic language. The classical Arabic is the language of Holy Quran, whereas the Modern Standard Arabic (MSA) is the language of books, news, official statements and magazines. On the other hand, the colloquial Arabic is most widely spoken and written in daily live conversations with different dialects. For example, Egyptian Arabic is spoken in Egypt. In GULF area, people speak Khaliji (GULF Arabic). Levantine Arabic" Shamii" is spoken by (Lebanon, Syria, Jordan). Maghrebi Arabic is spoken in Libya, Tunisia, Algeria, Morocco. In Sudan, people speak Sudanese Arabic. Omani Arabic is spoken in Oman. Yemeni Arabic is spoken in Yemen. And in Chad, people speak Chadian Arabic. The most used Arabic language over Social Media is the colloquial Arabic, although several users uses MSA or a mixture of MSA and colloquial Arabic.

Basically, In MSA language, every word has root. The rule of extracting roots is not an easy task. Every rule has many exceptions which create a new challenge in processing of MSA. One the other hand, Colloquial Arabic adds additional challenges. Each Arabic region has its own dialect and set of vocabularies. Moreover, words are formed without available dictionary for colloquial Arabic. Accordingly, there is no grammatic or formal structure in the colloquial Arabic. Additionally, several non-Arabic words are borrowed from other languages for example, the word "computer" is written in Arabic as (كمبيوتر) and pronounced as (Computer), and (to set up) is written as (يسطب) pronounced as (yesattab). The problem of such non-Arabic words doesn't have any known roots. So, there are no systematic methods can cover all these kinds of problems. The major researches in Arabic use root (stem) based methods for sentiment analysis process. However, the complexity of extracting the roots is not an easy task and gives low accuracy. So, working with words themselves without extracting the roots is that most suitable solution to overcome the previous challenges. Usually, any proposed solutions in sentiment analysis can be classified as Supervised, unsupervised, or hybrid techniques [8].

To this end, the main contribution in this paper is to propose a novel approach for Arabic sentiment analysis using a lexicon based approach. The proposed approach uses a hybrid technique for sentiment analysis. The hybrid approach is classified into supervised and unsupervised phases. The former phase extracts the polarity using a Look-up table stemming technique. In the later phase, the true classified polarity data, extracted from the first phase, is used to construct and train a classifier to classifies those data which are neither classified at all or false classified in the first phase. We test the performance of our proposed approach using MIKA corpus [1].

The rest of this paper is organized as follows. Section 2 gives background an overview on sentiment analysis. Section 3 presents related work. Section 4 describes the proposed approach. Section 5 discusses our experiments and evaluations. Section 6 concludes the paper.

## 2. Sentiment analysis

This section gives an overview on the sentiment analysis techniques. Since the paper focuses on Arabic sentiment analysis, this section starts with some basic characteristics of Arabic language.

### 2.1 Arabic language characteristics

Arabic language is one of six official languages used in the United Nations [5] and it is the native spoken language for more than 300 million people worldwide [9]. It is one of Semitic languages that is written from right to left and is consisted of 28 letters. The shapes of the written Arabic letters are changed according to their positions within the word. For instance, the letter (ج), pronounced as Jeem, is written as (جـ) if it comes at the beginning of a word, and it is written as (ج) if it comes in between the letters, whereas it is written as (ج) at the end of the word. In addition, Arabic consists of long and short vowels. The long vowels are denoted by letters, while the short vowels are not denoted as letters, but as diacritical marks instead. Each diacritical mark is written on either the top or the bottom of the letters. For example, the different diacriticals of the letter (ن), pronounced as (noon) can be (نَ or نُ, نْ). In fact, Arabic language is classified into three types; Classical, Modern Standard, and Colloquial Arabic language. The classical Arabic is the language of Holy Quran, whereas the Modern Standard Arabic (MSA) is the language of books, news, official statements and magazines. On the other hand, the colloquial Arabic is most widely spoken and written in daily live conversations with different dialects. For example, Egyptian Arabic is spoken in Egypt. In GULF area, people speak Khaliji (GULF Arabic). Levantine Arabic "Shamii" is spoken by (Lebanon, Syria, Jordan). Maghrebi Arabic is spoken in Libya, Tunisia, Algeria, Morocco. In Sudan, people speak Sudanese Arabic. Omani Arabic is spoken in Oman. Yemeni Arabic is spoken in Yemen. And in Chad, people speak Chadian Arabic.

Usually, each Arabic word consists of prefixes, root, and suffixes. The automatic extraction of the root of the word using a stemming technique is not an easy task, particularly in Colloquial Arabic. Colloquial Arabic is written without diacritical marks, which adds ambiguity to understand the meaning of the words. Take for example, the word (حب) without diacritic can be understood alone as (حُب) love, or (حَب) a seed.

### 2.2 Sentiment analysis techniques

Sentiment analysis [4] (opinion mining) identifies, extracts, and studies the attitude of the writer. The main task in sentiment analysis is classifying a document or a sentence based on a polarity. The polarity of document or sentence can be positive or negative. The existing techniques for sentiment analysis can be classified into three groups, unsupervised technique, supervised technique by means of machine learning, and hybrid technique [8]. The Unsupervised technique, also called lexicon based technique [10] usually builds a lexicon by containing roots or stems of words. Every root or stem associated with polarity (e.g. -1 for negative, +1 for positive). Basically, the Lexicon based technique consists of the following steps; dataset collection, tokenization, normalization, stop words removing, and stemming. Dataset collection means collecting the users' reviews or opinions from social media websites e.g. Facebook, Twitter, LinkedIn...etc. Tokenization is the process of splitting a text into set of separated word called tokens. In normalization, the common letters are put in the same form. For example, in Arabic the letter (أ) pronounced as "Alef" has four forms (ا, آ, إ, إ) is normalized into (ا). Also, words with unnecessary repeated letters are normalized into normal forms by removing the repetition. For example, the word "gooooooooood" will be normalized to "good". In Stop words removing, all stop words are removed from text. For example, stop words like; (و) translated as (and), (فوق) translated as (above) is removed. Finally, in the last Stemming step, the stem or the root of a word is extracted using a stemming technique. The root of word is unanalyzable morphology, but stem is consisting of the 'root' with a slight modification [11]. Generally, stemming techniques

can be classified into root-based, light, and lookup table stemming [12]. The root-based stemming technique [13], also called aggressive stemming, constructs a predefined list of all known roots, it removes the prefixes and suffixes from words and compares the result with the list of roots to find the exact root. The light Stemming, on the other hand, selects common prefixes and postfixes to truncate them from the words without dealing with infixes or finding the root of the word. For example: Larkey Stemmer [14] deletes the common used prefixes (e.g. ال, الا, بت, بي, بن, وال, و, ال, وا, فال) and suffixes (e.g. اخ, وا, ذا, ون, وه, ان, هه). The look-up table stemming technique [12] builds a table of words attached to roots. Then it searches by the word to find the exact root. For its main idea, it's the simplest stemming technique. The look-up table performance affected by the size of the table. As the number of words in the table increases, the performance gets higher. In any transactional system (e.g. Information Retrieval systems) look-up table will use a huge storage space as well as it'll consume a lot of time to search the look-up table. On the contrary, we believe it'll be efficient in Sentiment Analysis tools because the look-up table will contain only emotional words so it will not consume large space as well as, the process will not take long time. A lexicon of stems is constructed after finishing the previous steps. Each stem in the lexicon is attached to polarity. Considering an input dataset of reviews, each review is tokenized into set of tokens. The lexicon technique searches for the token in the lexicon to get the polarity. The polarity of an inputted review is calculated by summing all polarities of emotional tokens in the review.

In contrast to the lexicon based techniques, the supervised technique uses machine learning methods to extract the polarity [10]. It uses a training dataset (corpus of reviews annotated with negative or positive polarities) to extract dataset features (e.g. Ngram, POS (Part Of Speech), etc...). In this technique, a model is constructed using the extracted features based on one or more machine learning algorithms e.g. Support Vector Machine (SVM) [15] or Naïve Bays (NB) [16]. A dataset of reviews is used to test and evaluate the model based on the percentage of the true classified polarities. According to [17], the most widely used algorithms in sentiment analysis are SVM and NB. SVM is used in many Arabic sentiment analysis researches such as in [18], SVMlight is used in [19]. NB is used in many Arabic sentiment analysis researches such as in [20]. Both of SVM and NB are used in [21] the authors compared between the two algorithms.

Finally, the hybrid technique [10] combines both of the supervised and unsupervised techniques. The output of lexicon based technique is used as input for machine leaning algorithms, the true classified reviews used as a training set for the machine learning algorithm. After building the model using specific features. The unclassified reviews from supervised technique is used to test the model in the unsupervised technique.

### 3. Related work

Several researches have proposed solutions to sentiment analysis. In this section, we focus on those works that handle Arabic sentiment analysis. We classify those works according to the approaches that handle sentiment analysis. Several authors used the supervised machine learning approach to analyze sentiments. For example, in [18] the authors proposed a SVM-based classifier to classify user's comments into satisfied or unsatisfied comments. In their work, they used the light stemmer technique, proposed in [22] to extract the stem of words to build a lexicon of slang Arabic words and idioms (SSWIL). Another supervised work is proposed in [19]. The authors introduced SAMAR a machine learning system for subjectivity and sentiment analysis for Arabic social media genres. In their work, they applied an automatic tokenization tool proposed in [23] in which, they extracted the stem of words by isolating the prefixes and suffixes. They take into consideration several features (e.g. word form, Part of speech tagging). Finally, they applied SVMlight as the classifier in their experiments. An additional supervised technique is the work of [21]. The authors built their own dataset and crawled twitter for collecting tweets and they applied a Naïve Bayesian classifier to classify the input tweet into positive, negative or sarcastic.

They used language independent features (e.g. Ngram, sentence domain feature) and morphological feature. For experiments, they applied SVM on 80% of their data (training data) and the rest 20% as test set. Last but not least, a work proposed in [21] examined SVM and NB on Twitter data. They used Twitter API to collect their data. They used unigram and bigram as features to both SVM and NB. Their results showed that SVM is better than NB. In contrast to the work presented in this paper, the previous supervised methods suffer from drooping the accuracy. The highest accuracy achieved did not exceed 85%. Also, supervised techniques need a huge corpus to be annotated in order to train the model well.

Another group of works proposed to handle sentiment analysis using the unsupervised techniques, or the so-called Lexicon based techniques. For example, in [24] the authors worked in sentence level and document level to extract the sentiments of Arabic text. For sentence level approach, they applied a grammatical approach (Part of Speech) POS and a semantic orientation lexicon. They used the light stemming technique in the lexicon construction. On the document level, they used the sentiment of sentences to extract the document polarity. Another unsupervised work is presented in [25]. The authors built a tool for both MSA and colloquial Arabic comments and reviews. They built domain specific eight lexicons (e.g. Technology, Book, movies...). Moreover, they created two lexicons for every domain; one lexicon for positive polarity and the other one for negative polarity. They scaled polarities in the range from [0 to 10], where the range [0 to 4] represents weak polarity and [5 to 10] represents strong polarity. The scale was applied for both positive and negative lexicon. Their proposed approach can determine the domain of reviews, polarity (positive, negative, neutral), (subjective or objective), and (strong or weak). An addition unsupervised work is presented in [26]. The authors proposed a system to classify web-pages if containing reviews or not. Also, they built a sentimental analyzer to classify the business reviews into three classes (Positive, Negative, neutral or mixed). The classification is done base on their constructed lexicon of 600 positive words/phrase and more than 900 negative words/phrase and 100 neutral words collected from top frequent words/phrase over the Internet web-pages. Differently to the work presented in this paper, those unsupervised researches have some limitations. Firstly, the lexicon based technique cannot handle unseen words. If the stem of an emotional word is not found in the lexicon, the word will be neglected. Secondly, the accuracy is not still fairly high. The reason for that is due to the stemming problem. For example, the root based stemmer Khoja [13] has some defects while extracting the root of words. Take of example the word (حلو) sweet, pronounced as (helw), its correct root is (حلو) pronounced as (helw) but the extracted root from Khoja is (حلى) pronounced as (hala). Also, the stem-based algorithm Larkey [14] has some defects to extract the proper root. For instance, the word sweeter (أحلى) pronounced as (ahla), its correct stem is (حلو) pronounced as (helw), but it is extracted as (أحل) ahl.

Similarly, to the work proposed in this paper, several authors suggested a hybrid technique for sentiment analysis. For example, the author in [27], proposed a hybrid approach that combines a lexical based technique and the maximum entropy method followed by k-nearest machine learning method as the last phase. The author built their lexicon using SentiStrength project proposed in [28]. Also, the author translated the SentiStrength list into Arabic in association with word strength. Moreover, the author added some several Arabic words taken from on-line dictionaries. Another hybrid work is presented in [29]. The authors constructed a lexicon of MSA and Jordanian dialect. The lexicon was constructed with 300 and then was expanded to 3479 words consisting of 1262 positive words and 2217 negative ones. The authors used light stemming and root-based stemming approaches. They got 85% accuracy for root-based stemmer and 87% for light stemmer using SVM. Similar hybrid work is introduced in in [30]. The authors proposed a lexical based approach followed by SVM machine learning. They built their own lexicon of 1500 word classified into 1000 positive and 500 negatives. Additionally, they collected their dataset from Twitter and applied stop words removal and light stemming. Their results gave 67.6% accuracy on the lexicon based level and fed the result to SVM as the second phase. Ngram

technique was used as the feature of the machine learning model construction. The accuracy of second phase achieved 84%. In contrast to the work of this paper, we use look up table technique instead of the traditional stemming techniques with a bigger corpus.

#### 4. Proposed approach

In this section, we describe the proposed a hybrid approach for automatically extracting user's opinions from Arabic social media reviews. Our proposed approach combines the Lexicon method together with a supervised machine learning method to automatically analyze the input reviews into positive or negative. We begin this section by describing the overall architecture of our approach. Then we discuss the construction of the model in details.

##### 4.1 The general architecture

Basically, the proposed architecture consists of two major phases; the lexicon and the supervised learning phase. The main function of the lexicon phase is to extract the polarities (positive or negative) from a dataset  $D$  of unclassified reviews. The outcomes of this lexicon phase separate the dataset  $D$  of unclassified reviews into a set  $D_{classified}$  of true classified reviews, a set  $D_{unclassified}$  of unclassified reviews, and a set  $D_{false}$  of false classified reviews. The set  $D_{classified}$  is used to train a supervised classifier later in the supervised phase. In the later phase, both the set  $D_{unclassified}$  and  $D_{false}$  become a test for the classifier which is built using the set  $D_{classified}$ . The overall architecture is depicted in Figure 1.

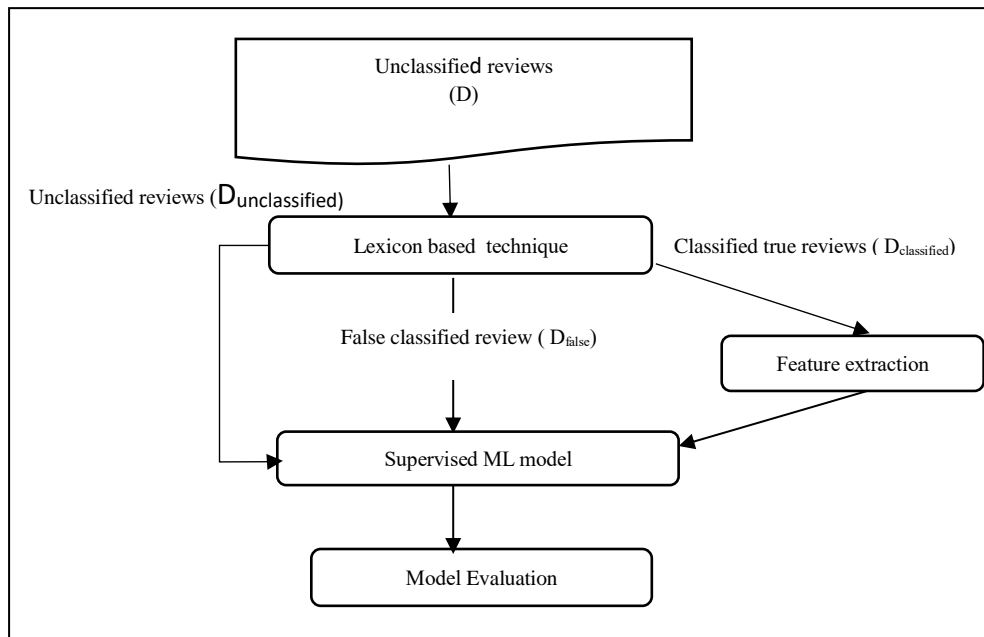


Figure 1. System architecture

##### 4.2 Model construction

In this subsection, we describe the steps of constructing our model. As mentioned previously, the proposed approach consists of two phases; the lexicon phase and supervised machine learning phase. in the following we show how to build the model in details in each phase.

###### 4.2.1 Lexicon based phase

This phase consists of multiple steps; the look-up table construction, Corpus (Dataset collection), and negation and emphasize list. To construct the look-up table, we build two tables; one for words, called look-up table, and another table for idioms. Our look-up table extends the proposed

lexicon in [31] for colloquial Egyptian Arabic and MSA sentiment words attached with their polarity.

The lexicon contains (989 idioms and 4974 Words). We extend the lexicon to construct a new Look-Up table for Arabic Sentiment Analysis with (7358 word) and a separated List of (1527 idioms). The look-up table, see Table 1, is constructed by inserting each word with an associated index, such that each word having possible variation forms are inserted with that same index. For example, the word (يُفسد) to corrupt, pronounced as (yafsad) has 55 possible different forms in Arabic language. All of these various form is inserted in the table with index 215.

Table 1. Look-up table structure

Index	Word
215	يُفسد (To corrupt)

This index in the look-up table is used to associate between the word and its root. Also, a lexicon of (552 root) is constructed, see table 2. Each root is associated with an index and polarity of the root. The look-up table and the lexicon is joined on the index. For example, the word (يُفسد) to corrupt in the look-up table has index 215 is joined with the lexicon on the same index to get the root (فسد) corrupted, pronounced as (fsad), and hence a negative polarity is retrieved. The steps of look-up table technique (Word-based technique) are; firstly, searching the word in the table. Secondly, get the word index.

Table 2. Lexicon Example

Index	Polarity	Word
215	-1	فسد (corrupted)

On the other hand, the table of idioms containing of 1527 Arabic Colloquial idioms is constructed. This table extends the idioms table existing in [31], by adding several idioms collected from tweets and online websites like [32]. Additionally, we add all variations of idioms that are most commonly wrongly written to cover the most typing mistakes. Each entry in the table of idioms contains the polarity of the idiom beside the idiom itself. Table 3 shows sample of idioms table.

Table 3. Idioms table

Expression	Polarity
(Less gentle) قليل الذوق	-1
(Nothing above it) لا يعلى عليه	1

For the dataset collection, we used the MIKA corpus proposed in [1]. The later corpus contains 2067 tweets on different topics e.g., reviews on Political, hotel reservation, products, and TV programs. It is classified into 1093 Positive tweets and 978 Negative tweets. The corpus is annotated as “PO” and “NG” representing Positive and Negative tweets respectively. Table 4 shows an example of the MIKA corpus.

Table 4. MIKA corpus example

Polarity	Tweets Example
Negative	NG=طبعا رامات 512 قليلة ... و المعالج من نوع غير جيد 512 rams are few, and the processor kid is not good = NG

Positive	PO=انا أدم أسامة الدلهماوى أميننا للقاهرة فى الإنتخابات الداخلية لحزب مصر القوية I support Osama Al-dlahmawe as Cairo secretary in the internal elections of the strong Egyptian party = PO
----------	--

In fact, there are some words that reverse the polarity of a word. For example: The word (لا) (don't), pronounced as (Laa), reverses the polarity of the word (أحب) (like) pronounced as (Auheb). When combined with the word (لا), it changes its polarity from positive to negative (لا أحب) (don't like). This is why we create a list of negation words. If the emotional word preceded by one of those negation words, the word polarity will be reversed. Examples of negation words (لا, لم, موش, مش, غير, لا, لم, موش, مش). They all give the meaning of (no, not, un...etc.). On the other hand, there are some words that strengthen the polarity of the emotional words. For example, the emphasize word (جدا very) duplicates the polarity of the positive word (جميل) (beautiful) when combined together. Similar, emphasize word (جدا very) duplicates the (وحش ugly) polarity of the negative word when combined together. In our setting, we handle this by creating a list of emphasize words. Examples of emphasize words include the words (قوى, جدا, فحت, طحن, كثير). They all give the different ways to say (too much, very much, very high, huge...etc.).

Now, it could be possible that a sentence contains a combinations of negation words and emphasize words associated with a sentiment. For example, if the emotional word preceded with a negation word and followed by an emphasize word. Firstly, the polarity of the word is reversed. Secondly the polarity is duplicated. For instance, the review (الفيلم مش كويس خالص) translated as (the movie is not good at all), the emotional word (كويس good) has +1 polarity, the negation word (مش not) polarity reverses the polarity to -1, and the emphasize word (خالص at all) duplicates the polarity to -2. Calculating the polarity of sentence is performed following the equation (1). In the later equation, S is used to get the sum of idioms' polarities in the tweet. B is used to calculate the polarity of first word in the review if it followed by emphasize word. So, the polarity of the word is duplicated. L is used to calculate the polarity of last word in the review if it preceded by negation letter. So, the polarity is reversed. P is used to calculate neither the first word nor the last word. If the word preceded by negation letter and followed by emphasize word too. The polarity of the word is reversed then duplicated. If the word is not preceded by negation letter or followed by emphasize word, then the polarity is the word's polarity without any changes.

$$S = \sum_{idioms} Polarity_{idioms} \quad B = (Polarity_{FirstToken} + EmphasisFactor_{FirstToken+1})$$

$$P = \sum_{token=2}^{n-1} [(Polarity_{token} \times NegationFactor_{token-1}) + EmphasisFactor_{token+1}]$$

$$L = (Polarity_{LastToken} \times NegationFactor_{LastToken-1}) \text{ where } EmphasisFactor \text{ default value equals zero, and } NegationFactor \text{ default value equals one.}$$

$$TweetPolarity = S + B + P + L \quad (1)$$

The previous equation (1) is translated into the following proposed lexicon algorithm, depicted in Figure 2, to extract the polarity. In Figure 2, steps 1 to 10, the algorithm searches for idioms in every tweet. If any idiom is found, the tweet polarity is updated with idiom's polarity. steps 11 to the end, the algorithm calculates the tweet polarity by calculating emotional words polarity according to (1). Figure 2 shows the lexicon based algorithm.

#### 4.2.2 Supervised machine learning

The limitation of the Lexicon phase is unseen words problem. If a tweet contains emotional word that doesn't exist in the look-up table, then the polarity of the whole tweet will not be correct. In this phase, we propose a solution of the unseen words problem by using a supervised machine



learning phase. In this phase, we use We used SVM [15] and NB [16] as a supervised learning technique, as they are the most common learning techniques in sentiment analysis according to [17]. Recall again, the lexicon phase is generally splits the Dataset D into correctly classified Dclassified, false classified Dfalse, and unclassified tweets Dunclassified.

<b>Input:</b> unclassified tweets.	<b>Output:</b> classified true tweets, unclassified tweets, and classified not true tweet.
<b>NegList:</b> a list of negation words.	<b>EmpList:</b> a list of emphasize words.
<b>ti-Prev:</b> the word before the token.	<b>ti-next:</b> the word after the token.
	<b>ti-pol:</b> polarity of token.
	<b>Tweet-pol:</b> polarity of tweet.

```

1. begin
2. for each tweet in corpus do
3. T ← tweet tokenized into list of tokens t1, t2, t3....
4. D ← List of idioms d1, d2, d3....
5. Tweet-pol ← 0
6. for each di in D do
7.   if di is found
8.     Tweet-pol ← Tweet-pol + Idiom_Pol
9.   end if
10. end for
11. for each ti in T do
12.   initialize Is-Reversed ← false
13.   Search for ti in the Look-Up Table
14.   if ti found
15.     return ti-pol in the lexicon
16.   end if
17.   if ti-Prev in the NegList
18.     Tweet_pol ← Tweet-pol + (ti-pol*-1)
19.     Is-Reversed ← true
20.   end if
21.   if ti-next in the EmpList
22.     if ti-pol is Positive
23.       if Is-Reversed ← true
24.         Tweet_pol ← Tweet-pol - 1
25.       else
26.         Tweet-pol ← Tweet-pol+ 1
27.       end if
28.     end if
29.     if ti-pol is Negative
30.       if Is-Reversed ← true
31.         Tweet_pol ← Tweet-pol + 1
32.       else
33.         Tweet-pol ← Tweet-pol - 1
34.       end if
35.     else
36.       Tweet-pol = Tweet-pol + ti-pol
37.     end if
38.   end if
39. end for
40. end for
41. end

```

Figure 2 The lexicon based proposed algorithm

In this machine learning phase, we build a classifier using Dclassified, to further classify Dunclassified and Dfalse.

This phase consists of multiple steps; pre-processing, feature extraction, model construction, and model testing step. In the pre-processing step, stop words are removed from training and test set. By stop words removal, we mean removing the common used stop words, which don't affect the meaning of the sentence in Arabic such as (من، فوق، على، و، أنا، شيء، هذه، هي، هو). Table 4 shows

an example of tweet after stop words removal. The stop words removal is applied in both training data and test data.

Table 4. Stop words removal example

<b>Before removal</b>	أفضل شي في التكييف أي شي صناعة تايلندي شارب ممتاز The best thing in the AC that was made in Thailand, Sharp is excellent.
<b>After removal</b>	أفضل التكييف صناعته تايلندي شارب ممتاز best AC made Thailand, Sharp excellent.
<b>Stop words</b>	شيء ، في ، أي Thing, in, the

In the feature extraction step, features are built and extracted by means of Ngrams. As we work with textual data, a feature vector [16] is created for each term of Ngrams. The vector that represents Ngrams and their frequencies in the dataset is  $\{(ti: fi), \text{tweet polarity}\}$ , where ti represents gram terms, fi represents the term frequency, and tweet polarity is the polarity of the whole tweet (-1 or 1). Table 5 Gives an example of tweet grams.

Table 5. Example of tweet grams

أفضل التكييف صناعته تايلندي شارب ممتاز best AC made Thailand, Sharp excellent.		
Unigrams	Bigrams	Trigrams
1. أفضل - Best	1. أفضل التكييف - Best AC	1. أفضل التكييف صناعته - Best AC made
2. التكييف - AC	2. التكييف صناعته - AC made	2. التكييف صناعته تايلندي - AC made Thailand
3. صناعته - Made	3. صناعته تايلندي - Made Thailand	3. صناعته تايلندي شارب - Made Thailand shard
4. تايلندي - Thailand	4. تايلندي شارب - Thailand sharp	4. تايلندي شارب ممتاز - Thailand sharp excellent
5. شارب - Sharp	5. شارب ممتاز - Sharp excellent	
6. ممتاز - excellent		

In the model construction, the classified true tweets  $D_{classified}$  is used as a training data for the supervised technique. In the testing step, the unclassified tweets  $D_{unclassified}$  and the false classified tweets  $D_{false}$  both are used as the test set for the constructed model.

## 5. Experimental results

This section presents the experimental results of the proposed approach. We experiment and evaluate our proposed approach using MIKA corpus [1]. We compare our result with the work of Khoja [13] and Larkey [14]. To study the impact of the proposed approach, we run three different experiments. The first experiment takes into the consideration the look-Up table without idioms. The second experiment considers the look-up table with enhanced list of idioms. Finally, the last experiment takes the look-Up table into consideration associated with the machine learning phase. Accuracy, Recall, Precision, and F-measure are taken into consideration as evaluation measures. Equation 2 to equation 5 defines these measures. The score of each measure is calculated based on TP, TN, FP, and FN, where TP (true positives) refers to the positive tweets that are classified correctly as positives, TN (true negatives) refers to the negative tweets that are classified correctly as negatives, FP (false positive) refers to the negative tweets that were classified incorrectly as positives, and FN (false negative) refers to the positive tweets that were classified incorrectly as negatives [30].

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$F = \frac{2TP}{2TP+FN+FP} \quad (5)$$

Table 6 shows the evaluation results of applying lexicon based approach using Look-Up table but without using idioms list. We compare the Look-Up stemming technique with light stemming technique (Larkey) and aggressive stemming technique (Khoja), but without taking into consideration the list of idioms in order to study the effect of using idioms on the evaluation measurements. Our proposed approach achieved accuracy of 82.58%, Larkey light-stemmer achieved accuracy of 70.88%, and Khoja aggressive- stemmer achieved accuracy of 58.59%.

Table 6. Lexicon based approach results

Stemming algorithm	Accuracy	Precision	Recall	F-Measure
Khoja	58.59%	63%	60.3%	61.7%
Larkey	70.88%	77%	70.4%	73.6%
Look-Up	82.58%	88.2%	80.7%	84.3%

Table 7 shows the evaluation of lexicon based approach using Look-Up table stemming technique while taking into consideration the list of idioms. Compared to Larkey stemmer and Khoja stemmer. Our proposed approach achieved accuracy of 90.86%, while Larkey light stemmer achieved 79.58%, and Khoja aggressive stemmer achieved 64.10%.

Table 7. Lexicon based approach results (idioms List)

Stemming algorithm	Accuracy	Precision	Recall	F-Measure
Khoja	64.1%	66.3%	65.9%	66.1%
Larkey	79.58%	81.9%	73.8%	80.8%
Look-Up	90.8%	92.8%	90.2%	91.5%

Further step of our experiments, we put the supervised machine learning technique into consideration to classify the unclassified tweets from the lexicons based phase. Here we train our model using SVM and NB. Both kinds of supervised techniques are trained using the true classified tweets coming from the lexicon phase. In particular, our results of the first phase show that the classified true tweets are (1878 tweet). The later Dclassified is used to train both SVM and NB. The unclassified tweets Dunclassified and the incorrectly/false classified tweets Dfalse both are used as the test set (189 tweet). SVM achieved 56.61% accuracy, 56.9% precision, 56.6

% recall, 56.7% f-measure. NB achieved 56.61% accuracy, 55.1% precision, 56.6% recall, 55.1% f-measure. Table 8 presents the evaluation result on the set Dunclassified and Dfalse using both SVM and NB.

Table 8. SVM and NB results using lexical output

ML Algorithm	Accuracy	Precision	Recall	F-Measure
SVM	56.61%	56.9%	56.6%	56.7%
NB	56.61%	55.1%	56.6%	55.1%

Combining the previous results to the results of the first phase, we find that applying the lexicon based approach together with machine learning approach achieved 96% accuracy. Table 9 summarizes the final result.

Table 9. Final results

Algorithm	Accuracy
Lexicon + Look-Up stemmer	82.58%
Lexicon + Look-Up stemmer + idioms list	90.8%
Lexicon + Look-Up stemmer + idioms list + SVM	96%

## Conclusion and future work

The wide spread of Social Media leads to a massive amount of data produced daily. Understanding and mining these data to determine the users' attitude towards products, services, events, and topics is very useful for individuals as well as for stakeholders. One of the popular contents sharing languages across social networks is the Arabic languages. The Arabic language has several challenges regarding finding the stem of a word especially in case of Colloquial Arabic. The later kind of Arabic language is the most widely spoken and written Arabic. In this paper, we proposed a new hybrid approach for Arabic sentiment analysis. In the proposed approach, we relieved some challenges of Arabic language by using look-up table stemming technique instead of traditional stemming techniques. The proposed approach consisted of two phases. The first phase the sentiment analysis is handled by means of unsupervised learning approach or the so-called lexicon based approach. To tune the accuracy, in the second phase, we used a supervised learning method to handle those unclassified or false classified sentiments resulted from the first phase. We used MIKA corpus [1] as a dataset of collected tweets. We extended the lexicon in [31] to construct our look-up table. We augmented our lexicon with list of idioms constructed using the existing idioms in the lexicon in [31] together with additionally collected idioms. We compared our proposed approach using look-up table stemmer with other two different stemming techniques; root based stemmer, and stem based stemmer. The experiments showed that our look-up stemmer is performing better than the both stemming techniques. Additionally, we proposed a second phase to overcome the unseen words problem of look-up table stemmer. We used the output of lexicon phase to train two classifiers; SVM and Naive Bays. The unclassified tweets and false classified tweets were used as a test set of the classifiers. We found that the performance was tuned and the accuracy increased. The overall performance of the proposed hybrid technique achieved 96% accuracy.

In the future work, we intend to extend our Look-Up table and lexicon to cover more Arabic colloquial dialects. In addition, we intend to find an effective solution for sarcastic reviews and the ambiguous reviews. We think that semantics and ontology may help to solve such problems.

### Acknowledgements

We thank Salma Amr El Werdany, Faculty of Dar Al-Oloom, Cairo University for her help in constructing the Arabic Lexicon.

### REFERENCES

- [1] H. S. Ibrahim, S. M. Abdou and M. Gheith, "MIKA: A Tagged corpus for Modern Standard Arabic and Colloquial Sentiment Analysis," in *In proceeding of the 2nd IEEE International Conference on Recent Trends in Information Systems (ReTIS-15)*, Kolkata, India, 2015.
- [2] A. Paul, "What is Web 2.0? Ideas, technologies and implications for education," JISC Technology and Standards Watch, Bristol, 2007.
- [3] R. Zafarani, M. A. Abbasi and H. Liu, *Social Media Mining An Introduction*, Egypt: Cambridge University Press, 2014.
- [4] B. Liu, *Sentiment Analysis and Opinion Mining*, China: Morgan & Claypool Publishers, May 2012.
- [5] U. N. G. o. E. O. G. Names, "UNGEGN list of country names," 2 May 2011. [Online]. Available: [https://unstats.un.org/unsd/geoinfo/UNGEGN/docs/26th-gegn-docs/WP/WP54\\_UNGEGN%20WG%20Country%20Names%20Document%202011.pdf](https://unstats.un.org/unsd/geoinfo/UNGEGN/docs/26th-gegn-docs/WP/WP54_UNGEGN%20WG%20Country%20Names%20Document%202011.pdf). [Accessed 17 July 2017].
- [6] H. Hazrat, "Arabic Language: Characteristics and Importance," *A journal of Humanities & Social Science*, vol. I, no. 2278-5264, pp. 11-16, 2013.
- [7] F. Salem, "Social Media and the Internet of Things - Towards Data-Driven Policymaking in the Arab World: Potential, Limits and Concerns," Mohammed Bin Rashid Al Maktoum Global Initiatives, Dubai, 2017.
- [8] M. Dholariya, A. Ganatra and D. Bhoi, "A Survey on Sentiment Analysis: Tools and Techniques," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 5, no. 3, pp. 6046-6050, 2017.
- [9] A. B. Al-Saleh and M. E. B. Menai, "Automatic Arabic text summarization: a survey," *Artificial Intelligence Review*, vol. 45, no. 2, p. 203-234, 2016.
- [10] W. Medhat, A. Hassan and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093-1113, 2014.
- [11] M. Aronoff, "Morphological stems: what William of Ockham really said," *Word Structure*, vol. 5, no. 1, pp. 28-51, 2012.
- [12] M. N. AL-Kabi and R. S. A.-. Mustafa, "Arabic Root Based Stemmer," in *The 2006 International Arab Conference on Information Technology*, Jordan, 2006.
- [13] s. Khoja, "Research Interests," Pacific University | 2043 College Way • Forest Grove, Oregon 97116, [Online]. Available: <http://zeus.cs.pacificu.edu/shereen/research.htm>. [Accessed 11 July 2017].
- [14] L. L. S. and C. M. E., "Arabic information retrieval at UMass in TREC-10," in *TREC-10 conference*, Gaithersburg, Maryland, 2001.

- [15] C. Vapnik and C. Vladimir, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, p. 273–297, 1995.
- [16] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, California: Prentice Hall/Pearson Education, 2003.
- [17] M. Ayyavaraiah, "Review of Machine Learning based Sentiment Analysis on Social Web Data," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 4, no. 6, pp. 2320 - 9801, 2016.
- [18] T. H. A. Soliman, M. A. M, A. R. Hedar and M. M. Doss, "MINING SOCIAL NETWORKS' ARABIC SLANG COMMENTS," in *IADIS European Conference on Data Mining*, Prague, Czech Republic, 2013.
- [19] M. Abdul-Mageed, M. Diab and S. Kübler, "SAMAR: A system for subjectivity and sentiment analysis for Arabic social media," *Computer Speech and Language*, vol. 28, no. 1, pp. 20-37, 2014.
- [20] M. Abdul-Mageed, M. T. Diab and M. Korayem, "Subjectivity and Sentiment Analysis of Modern Standard Arabic," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: shortpapers*, Portland, Oregon, 2011.
- [21] A. Shoukry and A. Rafea, "Sentence-level Arabic sentiment analysis," in *Collaboration Technologies and Systems (CTS)*, Denver, CO, USA, 2012.
- [22] S. R. Elbeltagy and A. Reafea, "An accuracy-enhanced light stemmer for Arabic text," in *ACM Transactions on Speech and Language Processing (TSLP)*, New York, NY, USA, 2011.
- [23] M. Diab, "Second generation AMIRA tools for Arabic processing: Fast and robust tokenization POS tagging, and base phrase chunking," in *the Second International Conference on Arabic*, Cairo, Egypt, 2009.
- [24] N. Farra, E. Challita, R. A. Assi and H. Hajj, "Sentence-level and Document-level Sentiment Mining for Arabic Texts," in *IEEE International Conference on Data Mining Workshops*, Beirut, Lebanon, 2010.
- [25] M. N. Al-Kabi, A. H. Gigieh, I. M. Alsmadi, H. A. Wahsheh and M. M. Haidar, "Opinion Mining and Analysis for Arabic Language," *International Journal of Advanced Computer Science and Applications*, vol. 5, no. 5, 2014.
- [26] M. Elhawary and M. Elfeky, "Mining Arabic Business Reviews," in *2010 IEEE International Conference on Data Mining Workshops*, Sydney, Australia, 2010.
- [27] A. El-Halees, "Arabic Opinion Mining Using Combined Classification Approach," in *Proceedings of the International Arab Conference on Information Technology*, Zarqa, Jordan, 2011.
- [28] M. Thelwall, K. Buckley, G. Paltoglou and D. Cai, "Sentiment strength Detection in short informal text," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 12, 2010.
- [29] N. Abdulla, N. A. Ahmed, M. A. Shehab and M. Al-Ayyoub, "Arabic Sentiment Analysis: Lexicon-based and Corpus-based," in *IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies*, Amman, 2013.
- [30] H. K. Aldayel and A. M. Azmi, "Arabic tweets sentiment analysis – a hybrid schema," *Journal of Information Science*, vol. 42, no. 6, p. 782–797, 2016.
- [31] El-Beltagy and S. R, "NileULex: A Phrase and Word Level Sentiment Lexicon for Egyptian and Modern Standard Arabic," in *In proceedings of LREC 2016*, Portorož, Slovenia, 2016.
- [32] H. Maxos, "The Arabic Idioms," in *The Art of Speech-Arabic Idioms*, Damascus, 2003.