Sentiment Analysis Challenges of Informal Arabic Language

Salihah AlOtaibi
Information Systems Department,
College of Computer and Information Sciences
Al Imam Mohammad Ibn Saud Islamic University (IMSIU)
Riyadh, KSA

Muhammad Badruddin Khan
Information Systems Department,
College of Computer and Information Sciences
Al Imam Mohammad Ibn Saud Islamic University (IMSIU)
Riyadh, KSA

***Abstract*—Recently, there are wide numbers of users that use the social network like Twitter, Facebook, MySpace to share various kinds of resources, express their opinions, thoughts, messages in real time. Thus, increase the amount of electronic content that generated by users. Sentiment analysis becomes a very interesting topic in research community. Thereby, we need to give more attention to Arabic sentiment analysis. This paper discusses the challenges and obstacles when analyze the sentiment analysis of informal Arabic, the social media. The most of recent research sentiment analysis conduct for English text. Also, when the research works in Arabic sentiment analysis, they focus in formal Arabic. However, most of social media network use the informal Arabic (colloquial) such as Twitter and YouTube website. This paper investigates the problems and the challenges to identify sentiment in informal Arabic language which is mostly used when users express their opinions and feelings in context of twitter and YouTube Arabic content.**

Keywords—Informal Arabic; Sentiment analysis; Opinion Mining (OM); Twitter; YouTube

I. Introduction

Arabic is a Semitic language spoken by more than 330 million people as a native language. Arabic is a highly structured and derivational language, in which morphology has a very important role. Thus, Arabic natural language processing (NLP) applications must deal with the complex nature of the Arabic language. For example, Arabic is written from right to left and no capitalization is used for nouns, which is a necessary feature in text mining. The Arabic language contains 28 letters and, in addition, the Hamza (ء). In Arabic, letters change their shape according to their position in the word (beginning, middle, or end) [1]. For example, see letter " ي/ya'a" and letter "ج/geem," as shown in Table I.

TABLE I. Position of the Character in the Word

Letter	Beginning	Middle	End
ي	يـ	ـيـ	ـي
ج	جـ	ـجـ	ـج

Arabic is the official language of Islam and of the last Prophet. It was selected to be the language of the Holy Qur'an. Muslims living throughout the world, thus, feel an affiliation with the Arabic language[1].

A. *Types of arabic*

There are three types of Arabic: Classical Arabic language (CA), Modern Standard Arabic language (MSA), and informal Arabic language (the latter is sometimes referred to as colloquial Arabic language).

CA is the language of Islam, which Arabic speakers use in their prayers and when reading the Qur'an. MSA is the official language across the Arab world. It is used by educated people in more formal circumstances; for example, for news reports, in classrooms, and in business. Informal Arabic is the language that people speak daily with family and friends, in which people also use their own dialects, which vary from region to region. The three different styles of Arabic language are available to every Arab – for example, each day, an Arabic speaker will use Classical Arabic for his five daily prayers, MSA when listening to or reading the news, and his/her own dialect when at home. Each type of Arabic has its own grammar, lexicon, and morphology, although though some properties are shared between the varieties. Most existing research tools have been developed to handle text that is written in MSA. This constitutes a limitation when it comes to research that focuses on text mining in relation to informal Arabic language [1], [2], [3].

In the research field, the sentiment analysis becomes hot topic to work in. The most of research and techniques for sentiment analysis is for English text. Thereby, it is obvious, there are limitations in the researches that interest for sentiment analysis for Arabic language [4]. Moreover, most of the researchers focus on formal Arabic language [5]. Since most of users use informal Arabic in the world of social media, the task of sentiment analysis becomes more sophisticated [6]. This motivates us to explore the challenges to analyze the sentiment for informal Arabic language such the different Arabic Dialects are another challenge.

The paper is organized in few sections to describe further details of our work. Section 2 describes the nature and the complexity of Arabic language. Section 3 gives overview about the main commons and differences between informal Arabic and informal English. Section 4 gives overview about the challenges in sentiment analysis for Arabic language. In section 5, outlines the related work done in this area. In section 6, gives overview about the main commons and differences between Twitter and YouTube dataset. In section 7, we describe the method and the preprocessing. Section 8 shows our finding and discussion. Finally, in the brief Section 9, we make concluding remarks.

II. THE COMPLEXITY OF ARABIC LANGUAGE

The Arabic language is challenging and complex due to its nature and characteristics. The following paragraphs illustrate the complexity of Arabic.

This section provides a literature review for the field of sentiment and semantic analysis, focusing mainly on informal Arabic language.

A. Word meaning

The term "word" defines a single, isolated item between two spaces, which has a certain meaning. In Arabic, it is common for one word to have several different meanings, depending on the context. Table II gives the example of the Arabic word سهل/sahel, which can be used as a noun with three different meanings. The phrases have been taken from Twitter [7].

TABLE II. MEANINGS OF THE WORD سهل/SAHEL AS A NOUN

Sentences	Phrases in English	Word Meaning
سهل منبسط *sahel moonbaset*	"flat plain"	Flat floor
سهل بن سعد *Sahel bin Saad*	"Sahel bin Saad"	Name
سَهْلُ المَرام *sahel almaraam*	"easy to get"	Easy

B. Variations in lexical category

In Arabic linguistics, a word can be a noun, verb, or particle. The term "particle" covers all other words that are not nouns or verbs, such as prepositions and conjunctions, for instance. Examples are given in Table III.

TABLE III. WORD TYPES IN THE ARABIC LANGUAGE

Word Type	Example	English Translation
Noun	كتاب	Book
Verb	يكتب	Write
Particle	على	On

Moreover, a word can belong to different lexical categories, depending on the context. Table IV shows how the word حلق/halq can be used in different parts of speech [7].

TABLE IV. LEXICAL CATEGORIES FOR THE WORD حلق/HALQ

Phrases	Phrases in English	Word Category	Word Meaning
حلَق الانسان *halq alensan*	"human throat"	Noun	Throat
حَلَق رأسه *halq ra'asah*	"shaving his head"	Verb	Shaving
حَلَق الطائر *halq alta'er*	"flying bird"	Verb	Fly

C. Morphological characteristics

Morphology is a branch of linguistics that deals with the structure of words. It concerns word formation, roots, and affixation behaviors. Arabic is a highly structured and derivational language. Arabic is a Semitic language and it is morphologically complex. Typically, a word in a Semitic language contains more information than a word in a non-Semitic language like English.

In Arabic, for example, various affixes can be attached to create new words; from the root word درس/*darasa*, for instance, several different words can be generated, such as يدرس/*yadras* ("studying" in English), مدرس/*modras* (English: "teacher"), مدرسة/madrasa (English: "school"), and مدارس/*madares* (English: "schools") [8]. Below is short description of each basic item in the Arabic language.

As clarified above, a word is a single, isolated item with a certain meaning. In Arabic, a word can be a noun, verb, or particle, and the same word can fit into different categories, depending on the context.

A morpheme is the smallest linguistic unit that has a meaning. A morpheme cannot be split into smaller units. Morphemes should give a meaning to the word of which they are a part.

A root is a single morpheme that provides the basic meaning of a word. In Arabic, the root is the original form of the word, before any transformation process occurs. Many words can be formed using one root.

A *stem* is a morpheme without an affix. The stem provides a specific idea or meaning. In English, the root is also sometimes called the "stem" or "word base," but in Arabic, the stem (or base) is different from the root [7]. Table V illustrates the morphological characteristics of Arabic.

TABLE V. MORPHOLOGICAL CHARACTERISTICS

Morphological characteristics	Definition	Example
Word	a single and isolated item between two spaces	المحمدون/alMuhammadwn
Morpheme	smallest linguistic unit that has a meaning	ون- / Wn
Stem	The basic form of word	محمد /Muhammad
Root	The original form of word	حمد/Hammd

An *affix* is a morpheme that can be added before (*prefix*), after (*suffix*), or within (*infix*) the root or stem to give a new word or meaning [7]. Table VI shows how the word سجد/ Sajed can have different meanings when various affixes are added.

TABLE VI. DIFFERENT MEANINGS OF سجد/ SAYED WHEN DIFFERENT AFFIXES ADDED

Word	English translation	Suffix	Infix	Prefix
ساجد/Sajeed	"Prostrate"	***	ا	***
مسجد/ Msadjad	"Msadjad"	***	***	م
سجادة/Sejada	"Carpet"	دة	ا	***

In text mining, the stemming process is usually used to convert a word into its root form. The main objective of the stemming process is to remove all possible affixes, thus diminishing the complexity of a word and reducing the number of features and tokens in corpora [7]. For example, if the words ذاهبون/*thahebon*, ذهبوا/*thahabo*, and يذهب/*yathhab* are all in a corpus, after the stemming process has taken place, all the words will be recognized in the text mining procedure as the same word ذهب/*thhab* ("go" in English). However, the stemming process is not always considered beneficial in

Arabic because the Arabic root is context dependent; thus, a stem may lead to more than one definition [9]. Table VII exemplifies words with different meanings that share a common root.

TABLE VII. DIFFERENT WORDS WITH THE SAME ROOT

Sentences	English translate	Root	Meaning
يخرج من المنزل *yakrooj men almanzel*	"leaves home"	خرج	Goes out
تخرج من الجامعه *takarraj men aljameea*	"graduates from college"	خرج	Graduate

Vowelization or *diacritization* is the process of putting diacritical mark vowels above or under letters in Arabic words (*fatha*: َ, *dammah*: ُ, *kasrah*: ِ). *Nunation* is the process of putting a set of diacritically marked vowels at the end of a word to create the sound of the letter ن/*N*. The *kasheeda* (**—**) or *tatweel* is the symbol used to stretch some Arabic characters [7]. The *tatweel* symbol is often used in informal Arabic language to emphasize a feeling or meaning. In the text mining process, the *tatweel* must be removed because it creates multiple forms of the same word. Table VIII shows how *tatweel* preformed different forms for one words.

TABLE VIII. TATWEEL

Word	English translation
مرحبا مرحبـا مرحبـــا مرحبـــــــا / marhaba	"hello"

III. THE INFORMAL ARABIC VS. INFORMAL ENGLISH LANGUAGE

Informal language could be described as language that ignores the standard rules of grammar and spelling. In general, the Arabic language is written from right to left, while English is written from left to right. There is no capitalization in Arabic, unlike in English [1].

Informal English uses abbreviations (for example, "m8" for "mate" and "u" for "you"), whereas in Arabic, there are no such abbreviations. In informal Arabic language, abbreviations called *Arabization* are used (like برب for "be right back" and لول for "laughing out loud"). Arabization is the process of translating new concepts and terminology into Arabic. In fact, with Arabization, users translate only the first letter of each word in the English phrase or sentence to create a new abbreviation in Arabic (so, using the previous example of "be right back," برب is "BRB"). The main commonalities between informal Arabic and informal English are the use of emoticons, texting-style abbreviations, and repeated letters or punctuation, which is added for emphasis [10].

IV. ARABIC SENTIMENT ANALYSIS CHALLENGES

NLP for Arabic is fraught with many challenges, some of which result from the structural and morphological complexity of the language. As mentioned previously, Arabic is a derivational language, which means that many words can be formed from three-letter roots. The resulting words may look similar, but have very different meanings. Arabic grammar is also highly complex, containing a variety of sentence structures, both verbal and nominal. A verbal sentence is one that starts with a verb phrase, whereas a nominal sentence starts with a noun phrase. Arabic also contains many word forms and diacritics [1], [4]. The complex features of the language make the task of analysis more difficult [11]. Furthermore, the semantic dictionaries or lexicons on offer for Arabic text analysis are limited. Indeed, future research should consider the necessity of creating morphological analysis tools for Arabic text analysis that can cover all word forms and can perform suffix, affix, prefix, and root extraction. Grammatical analyzers and/or part-of-speech (POS) taggers are also needed. Some morphological analyzers have been developed for use with the Arabic language, such as BAMA (the Buckwalter Arabic Morphological Analyzer) and MADA (the Morphological Analysis and Disambiguation for Arabic analyzer). There are no sophisticated POS taggers and lexicons tools in Arabic which identify all parts of speech and discover the difference of sentence's types. These issues present a challenge for sentiment mining, which generally requires both semantic analysis of words and grammatical analysis of text [4].

In fact, another major challenge that has surfaced due to the emergence of social media is that most of the Arabic language found on the internet is written in informal Arabic. The informal version of the language is unstructured in nature. Furthermore, many users utilize their own regional dialects, rather than opting for modern standard Arabic; for instance, the word شوف/*shoof*, which means "look" in English, might be used instead of the word أنظر/*onther*. Another important point is that informal Arabic does not use diacritics; thus, in some cases, the meaning of the word becomes ambiguous. For example, the words مُدَرّسة ("teacher") and مَدْرسَة ("school") look the same when written without diacritics (مدرسة). Social media has also given rise to the increased usage of letter repetition to emphasize the meaning or feeling associated with a word (شكراااااا – "thankssss," as opposed to شكرا – "thanks") [12].

Informal Arabic words usually do not have their own specific roots. Indeed, a stemmer will sometimes identify the same root for both the informal word and the formal word, as is the case with the terms راحه/*rahaah* (formal) ("comfort" in English), andنروح /*nrooha* (informal) ("go" in English), both of which take the root روح/*rooh* [13]. Another key trait in Arabic social media is the use of compound phrases and idioms to express opinions; e.g., يا ولد يا مطوع/*ya walad ya motaua* (a negative expression that belittles someone pretending to be religious). Compound phrases and idioms vary from one country to another. Also, that gives different sentiment polarities rather than its constituent words itself. According to previous examples, the sentiment polarity is negative while none of its constituent words are negative [14].

As most social media users utilize informal Arabic, the task of text analysis therefore becomes more challenging. The introduction of various dialects poses a further difficulty [6] as does the lack of literature on informal Arabic language [5]. These factors motivated us to focus on the problems that exist in informal Arabic, with the aim of encouraging more researchers to participate in this field.

V. Related Work

Sentiment analysis depends on using various techniques of machine learning, such as Knowledge-based, corpus-based, Naïve Bayes (NB), support vector machine (SVM) and maximum Entropy model (ME). Sentiment analysis can be applied on different types of content such as content of newspapers, review sites, tweets from twitter site [15].

A. Sentiment analysis on Arabic Content

The sentiment analysis for Arabic language became topic of interest for many researches to participate in this field. In one study, researchers presented an advanced technique for inferring sentiment orientation of social media sites focusing on the problems related to web dependent analysis [16]. New tool was developed that can be used for Arabic sentiment analysis. The proposed tool is divided into two techniques; NLP and human computation. The proposed system consists of two parts; game-based lexicon and sentiment analyzer parts. The first part is used to build the lexicon based on human computation, while the second part is a sentiment analyzer that takes each review and executes sentences segmentation [5].

Other researchers proposed a new technique for Sentiment Analysis and Subjectivity Analysis (SSA) for certain Arabic social media sites. Results demonstrated that the use of lexeme or lemma data is useful. On the other hand, there is a need for individualized solutions for every task and genre [8]. Also, there is research work performed to do the sentiment analysis for Arabic Facebook news pages. They used three machine learning classification techniques; Naive Bayes, SVM and decision tree are used to improve the sentiment analyzer [17]. Some researchers also, proposed a technique for extracting and analyzing Arabic business reviews that are available in forums and blogs. The system has two basic parts; reviews classifier and sentiment analyzer. First part classifies the web page. Second part for detecting the polarity of the sentences based on an Arabic lexicon [18]. In 2012 an advanced Arabic sentence level sentiment categorization technique was introduced that depends on two methods; a grammatical and semantic methods. [19].

B. Arabic Sentiment analysis on twitter

As we mentioned in previous paragraphs, the research on Arabic semantic is limited. One of those limited studies was provided by A. Shoukry and A. Rafea. They produce an application on Arabic sentiment analysis by classification the Arabic tweets. They used different ML classifiers and different features. They apply the SVM and naïve bayes and also try the combinations of classifiers [3]. Also, other researchers tried to find and explore the problems of sentiment analysis for informal Arabic. They apply their experiments on twitter. They use knowledge-based technique. There is a limitation in the number of Arabic sentiment lexicons, and the main challenge is to build lexicons for informal words [13].

VI. Twitter Data Vs. YouTube Data

Twitter is a microblog and social network that allows users to share their thoughts and express their opinions through short massages. While YouTube is a website designed for sharing video. In YouTube the users can restrict who views their videos with YouTube's privacy option. Also the users can post a comment and reviews on the videos that were viewing. There is some common and different between Twitter and YouTube Arabic text.

The most commons between Twitter and YouTube users' post are all of the users use informal language that ignores the standard rules of grammar and spelling. Also the posts contain emoticons, texting-style abbreviations, and repeated letters or punctuation added for emphasis.

The main differences, on Twitter, users produce short pieces of information known as "tweets" (limited to 140 characters). One can find a diverse range of topics within these tweets. Twitter users may post tweets expressing opinions about personalities, politicians, products, companies, and events, for instance [20], [21], [22]. Furthermore, some of the symbols used in tweets are language-independent. For example, "@" is utilized when users are referring to other users. "#" (hash tag) is used to mark topics or keywords—it is used to make messages more visible to other people. "RT" (re-tweet) is used when someone likes a tweet and wants to repeat it for their followers to see. The writing technique for tweets is fast and short. Users utilize acronyms and emoticons to express their opinions.

On YouTube, users produce reviews and opinions on contains of videos. There is no limited length for reviews posts. The posts only reviews or comment on contains of videos unlike the twitter tweets. There are no special symbols used in reviews like tweets.

VII. Method

This paper aims to investigate the problem and challenges of informal Arabic sentiment analysis. In this paper, we used twitter and YouTube datasets. The processing of the method can be described as follows: 1) after collecting the datasets, we determine the annotation of each tweets and each YouTube review (positive, negative, and neutral). 2) Convert the emotion icons to text. 3) Clean the dataset by removing: names, URL, pictures, English word, for tweets re-tweets sign, hash tags. 4) Normalizing process which makes the text in consistent form, in other words, convert all different forms of word to a common form. 5) Tokenization process applied on each tweets to divide them into multiple tokens based on whitespaces characters. 6) Then make stemming process to return each word to its root. 7) Remove the Arabic stop-word. The result of preprocess is used as input to the classifier model to test the result. The sentiment classifier used in the model is Naïve Bayes algorithm.

VIII. Findings and Discussion

Informal Arabic language, in general, is "noisy" and poorly structured. It also features the non-standard repetition of letters, abbreviations, and emoticons, as well as the use of Arabized words.

Arabic tweets and YouTube reviews contain incorrect and misspelled word(s). These spelling problems needs special attention and require proper cleaning. When applying sentiment analysis for informal Arabic many problems occurred in text processing step. There are various problems that were found in each text processing phase. The following sub-sections expound the problems in each phase:

A. *Tokenization phase*

When applying sentiment analysis for informal Arabic many problems were encountered. The problems explained as following

1) Repetition Letters

The first problem is the repetition of letters, as mentioned in section 4. As we know that in the Arabic language if we have repeated letters in the text it cannot occur more than twice. So if the repetition exists at beginning, middle or at the end of the word more than two times, it will be detected in the pre-processing step. Unfortunately, repetition cannot be detected where a letter is repeated only twice. Table IX shows pre-processing of tweets with repetition letters. In literature issue of detection of the repetition is discussed for situation with repetition only existing at the end of word [13].

TABLE IX. REPETITION LETTER PROBLEM

Platform	Sentences	English Translate	After pre-Processing
Twitter	كئيب ل ابعععععد حد Kaeeb le abeeed haad	I am very depressed	كئيب ل ابععد حد
Twitter	هههههههه جميل جداً :) hahahahah Jamel jeedan :)	Very beautiful :)	هه جميل جدا :)
YouTube	أحسسسسسسسسسسسسسسسن Ahsssssssan	Better	أحسسن
YouTube	بالصراحة مرة واووووو beSaraha Marra wowwwww	In fact it is wow	بالصراحة مرة واوو

2) Negations

The second problem is that word polarities are affected significantly by ignoring negations like ما/*Ma*, لا/*Laa*, لم/*lam*, and لن/*lan* which are formal Arabic negations. The informal Arabic contains many of informal negation words like مو/*Muo*, مش/*Mush*, and موب/*Moub*, which also affect the text polarities by converting the meaning of the sentence to exactly the opposite. Furthermore, as we mentioned in section 3, the informal Arabic used Arabized words. The Arabized words "نو" and "نوت" which means in English "no" and "not", are also used as negations words in informal Arabic. Table X shows how the informal negation words affected the text polarities.

A negation indicator should, therefore, be used to detect polarities accurately.

TABLE X. WHO NEGATIONS AFFECT THE TEXT POLARITIES

Platform	Sentences	English Translate	Polarities	Sentences Without Negation	Polarities
Twitter	**مش** أهبل :) Mush Ahbal	Not idiot	positive	أهبل :)	negative
Twitter	ليش انا **مو** جريئه :(Leash Ina mu jareeah :(	Why I am not bold	negative	ليش انا جريئه :(	positive
Twitter	**نوو** يفهم شي No yefham shee	Does not understand something	negative	يفهم شي	positive
YouTube	**مو** ملاهي ذيي Mu malahe thee	This is not amusement park	negative	ملاهي ذيي	positive
YouTube	انا **موب** طفل Ana moub teffel	I am not a child,	Positive	انا طفل انا 14 سنه	negative

3) Connecting different words together

The third problem involves Twitter users connecting different words together—this method of writing occurs frequently in tweets because the length of a tweet is limited. This issue affects stop-word filtering because certain stop words are not removed and new forms of words are created. Table XI illustrate how this problem affects the pre-processing step by increasing the number of tokens

TABLE XI. THE EFFECT OF CONNECTING DIFFERENT WORDS TOGETHER AT TOKENIZATION AND STOP WORDS FALTERING

Tweet Problem	Sentences	English Translate	Tokenizing Process	Falter Stop Word
Tweet contain connecting words together	يامرحبا تسلم عزيز ابوخالد Yaa marhaba teslam aziz ibu Kaled	Hello, thank you Ibu Kaled, you are dear and precious person	يامرحبا تسلم عزيز وغالي ابوخالد	يامرحبا تسلم عزيز وغالي ابوخالد
Tweet does not contain connecting words together	يا مرحبا تسلم عزيز ابو خالد Yaa marhaba teslam aziz wa ghali ibu Kaled		يا مرحبا تسلم عزيز ابو خالد	مرحبا تسلم عزيز خالد

From the table above, shows the results of tokenization process and faltering the stop words are different based on how the tweet is written.

Connecting different words together can also cause ambiguities in meaning like words وفي/*wafee* and وهم/*whum* have two different meanings with/without connection as can be seen in Table XII.

TABLE XII. THE CONNECTING DIFFERENT WORDS TOGETHER CAUSE AMBIGUITIES IN MEANING

Platform	Sentences	English Translate	Word	Meaning
Twitter	تبتسم :)) . وفي عينيك ألف دمعة Tebtassen wafee Eaneek alf damaah	Smiling :)). and Thousands of tears in your eyes	و+في	And in
Twitter	قلبي وفي . ماني مثل غيري Galbe wafee mane methel garre	I have the loyal heart .. I am not like the other	وفي	Loyal
Twitter	الناس يغلطون وهم اللي يزعلون 🔪😒 Nass Agtaiwn waahum elle yezalon	people make mistakes and also they Angry	و+هم	And they
Twitter	وهم و حيره Waham waa heera	Illusion and confusion	وهم	worry
YouTube	وهم مايعترفوا بفشلهم Waa hom Mayereefo be fashalho	They did not admit for they failing	و+هم	And they
YouTube	وفي مقاطع فيديو رعب Waa fee makateea video roob	And in video clips horror	و+في	And in

4) Diacritization problem

The tokenization is performed based on finding whitespaces characters. Some types of punctuations like diacritic are removed and then add single space, so the word broken to many tokens. The problem was variations of word forms and diacritic. Table XIII shows the diacritic problems.

TABLE XIII. THE DIACRITIC PROBLEMS DURING TOKENIZATION PROCESS

Platform	Tweet before Tokenization	Tweet after Tokenization
Twitter	إن الصَلاةَ كَانت عَلَى الْمؤْمِنينَ كِتاباً موقوتَا "ena alsalat kanat ala almoemenen ketaaban moqouta"	إن الص لاة ك انت ع ل ى ال مؤ م نين ك تابا موقوت
Twitter	راح انْتَحَر ساعدوني "Raah Inteheer saodony"	راح ان ت حـ ر ساعدوني

The problem of the deletion of diacritics and certain word forms, like tatweel cases, was discussed in section 2. The problem was solved in this study during the suggestion pre-processing stage. Table XIV shows the normalization cases that were used in pre-processing.

TABLE XIV. NORMALIZATION CASES

Rule	Example
Tashkeel	المؤمنين<-الْمؤْمِنينَ
Tatweel	الله<-اللـــــه
Alef	آor أor إ->ا
Heh	هor ة-> ه

5) Emoticons problem

Informal Arabic language text often uses emoticons, which cannot be interpreted by text-based models.

When the text was filtered to remove English words and special characters, all the emoticons were also removed. Thus, to preserve the emoticons, meaningful names were given to each symbol appearing in the corpus, which allowed the role of emoticons to be examined at sentiment analysis model. Table XV shows examples of the emotion icons conversion step.

TABLE XV. EXAMPLES OF THE CONVERTING EMOTION ICONS TO MEANINGFUL TEXT

Platform	Emotion icons	Sentences	English Translate	After converting the icons
Twitter	♥)':	مع السلامه :') ♥ "Maa alslamah ♥)':"	Goodbye)': ♥	مع السلامه رمزحزين رمزقلب
Twitter	O_o	متردد O_o "Motaraded O_o"	Hesitant O_o	متردد رمزمتفاجئ
YouTube	(:	حلوه وتضحك (:هههههه..	Sweet and laugh (:	حلوه وتضحك هههههه.. رمزمبتسم
YouTube	._.	غبي ._.	Dumbass ._.	غبي رمزمتفاجئ

6) Writing style in informal Arabic text

Some writing styles used in informal Arabic text can affect text pre-processing results, such as when a word is written inside another word, or write the word in separate letters to emphasize the meaning or feeling, as shown in Table XVI.

TABLE XVI. EXAMPLES OF WRITTEN STYLES USED IN INFORMAL ARABIC LANGUAGE, AND TOKENIZATION PROCESSING RESULTS

Problem	English Transla-tion	Formal Sentence Style	Tokeniz-ation Processin g Results	Informal Sentence Style	Tokeniz -ation Processi ng Results
Writing a word inside another word.	Welcome	اهلا و سهلا *Ahlan wa sahlam*	اهلا و سهلا	اهـ وسهلا ـلا *Ahlan wa sahlam*	اهـ وسهلا ـــلا
	Hello or *Aslam alukom*	السلام عليكم *Aslam alukom*	السلام عليكم	السـ عليكم ـلام *Aslam alukom*	السـ عليكم ـلام
Word with separate letters	I finished	أنا منتهية *Ana Mentahea*	أنا منتهية	أنا م ن ه ي ه *Ana Mentahea*	أنا م ن ه ي ه

B. *Filter Arabic stop words phase:*

There is no given stop word list for informal Arabic language which contain informal Arabic words like: هاذي/*hathe*, هاذا/*hatha*, دي /*dee*, اللي/*elle*, so we build our own stop word list for informal Arabic language.

C. *Stemmer phase:*

In the Arabic there are different words with different meaning have the same root. This makes detecting the polarities of these words incorrect. As we mentioned above, in section 3.

Also other problem occurs during the stemming process. The stemmer some time deleted some basic letters the word Table XVII shows the light stemmer problems. We remove the stemmer step from the text processing.

TABLE XVII. Stemmer Deleted Some Basic Letter from the Word

Platform	Sentences	English Translate	Stemmer results
Twitter	القران الكريم❤ *AL Quran al Kareem*	Koran Kareem ❤	القران → قر
Twitter	انزين انتهينا *Enzaeen entahena*	now we finished	انزين → انز
Twitter	يا الله ساعدني *Ya Allah saedney*	O God, help me	الله → له
YouTube	يا الله مقرف *Ya allah mogreef*	O God, disgusting	الله → له
YouTube	هذا فلم كرتون *Hatha film carton*	This film carton	كرتون → كرى

IX. Conclusion

The Arabic language is both challenging due to its complex linguistic structure and interesting because of its history and importance in religion, culture, and literature. Informal Arabic language, in general, is "noisy" and poorly structured. It also features the non-standard repetition of letters, abbreviations, and emoticons, as well as the use of Arabized words. Thus, these features should be considered during text mining. This paper investigates the problems and the challenges to identify sentiment in informal Arabic language in context of twitter and YouTube Arabic content. In this experiment, we found many issues that can be motivating for future research

References

[1] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions," *ACM Trans. Asian Lang. Inf. Process.*, vol. 8, no. 4, p. 14, 2009.

[2] M. Korayem, D. Crandall, and M. Abdul-Mageed, "Subjectivity and sentiment analysis of Arabic: A survey," in *Advanced machine learning technologies and applications*, vol. 322, Berlin & Heidelberg, Germany: Springer, 2012, pp. 128–139.

[3] H. Froud, A. Lachkar, and S. A. Ouatik, "Arabic text summarization based on latent semantic analysis to enhance Arabic documents clustering," *Int. J. Data Min. Knowl. Manag. Process*, vol. 3, no. 1, pp. 79–95, 2013.

[4] N. Farra, E. Challita, R. A. Assi, and H. Hajj, "Sentence-level and document-level sentiment mining for Arabic texts," in *2010 IEEE international conference on data mining workshops (ICDMW), 13 Dec. 2010*, 2010, pp. 1114–1119.

[5] A. A. Al-Subaihin, H. S. Al-Khalifa, and A. S. Al-Salman, "A proposed sentiment analysis tool for modern Arabic using human-based computing," in *Proceedings of the 13th international conference on information integration and web-based applications and services*, 2011, pp. 543–546.

[6] A. Shoukry and A. Rafea, "Sentence-level Arabic sentiment analysis," in *2012 international conference on collaboration technologies and systems (CTS), 21-25 May 2012*, 2012, pp. 546–550.

[7] I. A. Al-Sughaiyer and I. A. Al-Kharashi, "Arabic morphological analysis techniques: A comprehensive survey," *J. Am. Soc. Inf. Sci. Technol.*, vol. 55, no. 3, pp. 189–213, 2004.

[8] M. Abdul-Mageed, M. Diab, and S. Kübler, "SAMAR: Subjectivity and sentiment analysis for Arabic social media," *Comput. Speech Lang.*, vol. 28, no. 1, pp. 20–37, 2014.

[9] A. Moh'd Mesleh, "Support vector machines based Arabic language text classification system: Feature selection comparative study," in *Advances in computer and information sciences and engineering*, T. Sobh, Ed. Netherlands: Springer, 2008, pp. 11–16.

[10] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," *J. Am. Soc. Inf. Sci. Technol.*, vol. 61, no. 12, pp. 2544–2558, 2010.

[11] P. Pak, Alexander and Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," *Lrec*, vol. 10, pp. pp. 1320–1326, 2010.

[12] R. M. Duwairi, R. Marji, N. Sha'ban, and S. Rushaidat, "Sentiment Analysis in Arabic tweets," in *5th international conference on information and communication systems (ICICS), 1-3 April 2014*, 2014, pp. 1–6.

[13] L. Albraheem and H. S. Al-Khalifa, "Exploring the problems of sentiment analysis in informal Arabic," in *Proceedings of the 14th international conference on information integration and web-based applications and services*, 2012, pp. 415–418.

[14] S. R. El-Beltagy and A. Ali, "Open issues in the sentiment analysis of arabic social media: A case study," in *2013 9th international conference on innovations in Information Technology (IIT), 17-19 March 2013*, 2013, pp. 215–220.

[15] A. Kumar and T. M. Sebastian, "Sentiment analysis on Twitter," *Int. J. Comput. Sci. Issues*, vol. 9, no. 4, pp. 372–378, 2012.

[16] R. Colbaugh and K. Glass, "Estimating sentiment orientation in social media for intelligence monitoring and analysis.," in *2010 IEEE international conference on intelligence and security informatics (ISI)*, 2010, pp. 135–137.

[17] A. E.-D. A. Hamouda and F. E. El-Taher, "Sentiment analyzer for Arabic Comments System," *Int. J. Adv. Comput. Sci. Appl.*, vol. 4, no. 3, pp. 99–103, 2013.

[18] M. Elhawary and M. Elfeky, "Mining Arabic business reviews," in *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, 2010, pp. 1108–1113.

[19] M. Abdul-Mageed and M. T. Diab, "AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis.," in *LREC*, 2012, pp. 3907–3914.

[20] K. Makice, *Twitter API: Up and running: Learn how to build applications with the Twitter API*. O'Reilly Media, Inc., 2009.

[21] L. Barbosa and J. Feng, "Robust sentiment detection on Twitter from biased and noisy data," in *Proceedings of the 23rd international conference on computational linguistics: Posters*, 2010, pp. 36–44.

[22] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Proj. Report, Stanford*, pp. 1–6, 2009.