PREPROCESSING ARABIC DIALECT FOR SENTIMENT MINING: STATE OF ART

Zineb NASSR[1], Nawal SAEL [2], Faouzia BENABBOU [3]

[123] Laboratory of Modelling and Information Technology Faculty of Sciences Ben M'SIK University Hassan II, Casablanca, Morocco
[1] nassrzineb@gmail.com, [2] Saelnawal@hotmail.com, [3] Faouzia.benabbou@univh2c.ma

KEY WORDS: Pre-processing, Arabic Dialect, Sentiment mining, Stop-words, Stemming, Lemmatization, Tokenization

ABSTRACT:

Sentiment Analysis concerns the analysis of ideas, emotions, evaluations, values, attitudes and feelings about products, services, companies, individuals, tasks, events, titles and their characteristics. With the increase in applications on the Internet and social networks, Sentiment Analysis has become more crucial in the field of text mining research and has since been used to explore users' opinions on various products or topics discussed on the Internet. Developments in the fields of Natural Language Processing and Computational Linguistics have contributed positively to Sentiment Analysis studies, especially for sentiments written in non-structured or semi-structured languages. In this paper, we present a literature review on the pre-processing task on the field of sentiment analysis and an analytical and comparative study of different researches conducted in Arabic social networks. This study allowed as concluding that several works have dealt with the generation of stop words dictionary. In this context, two approaches are adopted: first, the manual one, which gives rise to a limited list, and second, the automatic, where the list of stop words is extracted from social networks based on defined rules. For stemming two, algorithms have been proposed to isolate prefixes and suffixes from words in dialects. However, few works have been interested in dialects directly without translation. The Moroccan dialect in particular is considered as the 5th dialect studied among Arabic dialects after Jordanian, Egyptian, Tunisian and Algerian dialects. Despite the significant lack in studies carried out on Arabic dialects, we were able to extract several conclusions about the difficulties and challenges encountered through this comparative study, as well as the possible ways and tracks to study in any dialects sentiment analysis pre-processing solution.

1. INTRODUCTION

Nowadays, social networking has become in some ways one of the most popular communication tools. Social network environments are used by people of all ages, cultures and social categories to convey variant messages and can reach a global audience. Several platforms on the Web and social networks like Facebook, Twitter… allow peoples to convey opinions, share experiences or simply talk about everything about them online (Tan et al., 2011).The monitoring of social media has become an important way to analyse and detect trends, by studying and evaluating opinions on various topics such as politics (Eason et al., 1995), services (teachings, health...), marketing and business products.

People can share their opinions in an environment without constraint and, companies can extract useful ideas for their decision-making process. To quantify what individuals think from textual qualitative data, a polarity classification task for detecting positive, negative, or neutral text is required. Although, there is a large amount of research available on the analysis of documents such as newspaper, articles, journals, there are still several open questions to tackle concerning the real nature of the messages available online social networks.

Actually, many works are devoted to sentiment analysis from textual data over structured languages. However, much less effort is committed to providing a precise classification of sentiment for unstructured languages in general and more specifically for the Moroccan dialect "Darija". In our last work (Nassr et al., 2019; Nassr et al., 2019) we carried out a state of art and a comparative study of researches done in recent years on sentiment analysis. In addition to other findings, we deduced that, most of researches translate the comments in a structured language such as English to analyse them and that there are no standard resources for unstructured languages, such as Moroccan Darija (MDL). Given the fact that comments in Darija can be written in Arabic characters, Latin characters or a mixture of the two, this makes automatic processing more difficult to achieve.

In fact, user-generated content on the Web is generally unstructured and needs important pre-processing steps and analysis to extract useful knowledge (Melville et al., 2019). These steps depend on the nature of the language (structured or unstructured) and generally are different from one research to another. Their objective is to clean, normalize, transform and reduce the data size in order to adapt it to the learning algorithm.

The complete process of sentiment analysis includes data collection steps, pre-processing of the text, sensing of sentiment and its classification. Nevertheless, the pre-processing step is the most important in the analysis of feelings because messages in the social networks are characterized by colloquial expressions, abbreviations, emoticons, a lengthening of words, irregular capital letters, and do not generally conform to the canonical grammatical rules.

The objective of this work is to develop a comparative and statistical study of researches in Arabic Dialect for sentiment analysis. The paper is organized as follows: Section II develops the sentiment analysis background. Section III presents the related works. Section IV discusses our comparative study and the next one exposes the statistical analysis. Finally, the conclusion and the future works are detailed in section VI.

2. PREPROCESSING BACKGROUND

2.1 Pre-processing task Major steps

The pre-processing is a very important phase in sentiment analysis. It allows for quality data to be obtained and ensures a better performance in this analysis. This phase can be carried out in several stages, which depend on the nature of the language and the analysis objectives. The main stages are:

Data cleaning: one of the most important tasks in successfully mining social networks is cleaning up noisy data (Gharatkar et al., 2017).

Stop words removal: activity for removing words that are used for structuring language but do not contribute in any way to its content. Some of these words are a, are, the, was…

Tokenization: task for separating the full text string into a list of separate words. This is simple to perform in space-delimited languages such as English, Spanish or French, but becomes considerably more difficult in languages where words are not delimited by spaces like in Japanese, Chinese and Thai.

Stemming: heuristic process for deleting word affixes and leaving them in an invariant canonical form or "stem". For instance, person, person's, personify and personification become person when stemmed. The most popular English stemmer algorithm is Porter's stemmer.

Lemmatization: algorithmic process that brings a word into its non-inflected dictionary form. It is analogous to stemming but is achieved through a more rigorous set of steps that incorporate the morphological analysis of each word.

Despite the identification of these different stages, pre-processing phase is confronted by several problems, which are related to the sentiment analysis context. Indeed, words belonging to different parts of speech must be treated according to their linguistic role (adjective, nouns, verbs, etc.). The word style (bold, italic and underline) is not always available on online social media platforms and is often replaced by some language conventions. The lengthening of words like "it's seeeeeerious" (commonly known as expressive elongation or word stretching) is an example of new language conventions that are today very popular on online platforms. Other problems are related to additional terms such as the abbreviation expressions that are additional paralinguistic elements used in non-verbal communication (Lui et al., 2012) on online social networking platforms. The hashtags, which are widely used on online social networks to express one or more specific feelings. The distinction between sentiment hashtags and subject hashtags is a challenge that must be properly addressed for polarity classification and the emoticons, which are introduced as non-verbal expressive components in the written language to reflect the role played by facial expressions in speech.

Another very important pre-processing challenge is having to detect and analyze the uppercase letters given that positive and negative expressions are commonly reported by the uppercase of certain specific words (for example, ' #StarWars was UNBELIEVABLE! ') to express the intensity of the user's feelings.

2.2 Arabic dialect pre-processing challenges

The pre-processing phase is faced by several problems and challenges. These challenges are more important in the case of feelings written in unstructured languages. Sentiment analysis for Arabic dialects in general and Moroccan Darija in particular suffers from several complex problems, related to its nature, such as:

• Replacement of the kasra ("i" vowel/sound as in liberty) by the sukun (diacritic that marks the absence of a vowel) at the beginning of a word, as in ("كتاب ktAb") instead of ("كتاب kitAb") in Modern Standard Arabic (MSA);

• Bypassing or avoiding the Hamza to be pronounced as a "ya" sound, e.g. "3A'ilah" (family) is uttered "3Aylah";

• Some pronouns are slightly modified from their MSA form. E.g. ("أنتم antoum") for ("نتوما ntouma"), (« نتي ntiya˚ » or «نت nti˚») instead of (« انت anti ») ;

• For the possessive, it is common to add the word (ديالي my) instead of just the MSA suffixed pronoun. For instance, one would equivalently say (« كتابي ktAby") or (الكتاب ديالي al-ktAb dyali) to say « my book » ;

• Some of the interrogative particles are slightly modified. E.g. ("فين wyn" for "where"), ("شكون shkoun" for "who").

• The negation is introduced by means of the word (« ما mA ») and the suffix ْش with the sukun on it (« sh ») as in (ماكليتش mA rAnysh, "I am not") or (ماراناش ما mA rAnAsh,"We are not")… .Negation also has some other expressions such as (مارانيش

• As mentioned above, a number of words that are not of Arabic origin have percolated into Moroccan dialect, such as (طابلة "TAblah") for Table from French,(كارطابلي ") Kartably") for "my schoolbag", from French...

• Most often, words of non-Arab origin are conjugated using the rules applied to those of Arabic origin. For instance,(طابلة ") TAblah" for Table) gets a plural form as ("طبلات TAblAt" for Tables) following the regular plural of Arabic feminine nouns (which the form used for all nouns of foreign origin) (Al ayyoub et al., 2019). Verbs of foreign origin are also conjugated as if they were of Arabic origin. For instance, ("مافرناتش ma franatch", "she did not pull the brake").

3. RELATED WORK

Several studies have been interested in sentiment analysis and a variety of approaches have been developed, especially for English language. However, research studies are more limited when it comes to other languages, such as Arabic. This section discusses research studies in the field of sentiment analysis for Arabic dialects.

Stops word detection and elimination is one of the major challenges of sentiment analysis in the context of non-structured languages. In MSA and Arabic Dialects, there is no general standard stop-word list to use. To overcome this lack, for instance, (Walaa et al., 2015) generated stop words list from Online Social Network (OSN) corpora like Twitter, Facebook…etc for Egyptian Dialect (ED) The methodology consists of three phases: calculating the words' frequency of occurrence, checking the validity of a word to be a stop word, and adding all possible prefixes and suffixes to the words generated. (Alajmi et al., 2012) generated Arabic language stop words list. The list generation involves various important factors like word frequency calculation, mean and variance calculation, Entropy calculation, and Borda's ranking. (Khouja et al., 1999) created her Arabic stemmer with 168 words; this has been used by (Larkey et al., 2001; Larkey et al., 2002). A top-list minted through translating an English list and enhancing it with high frequency words from the corpus leading to a larger

1.131-word list has been developed by Chen and Gey., 2001). Moreover, a dependent domain list, which includes three problems, has been created by (Savoy et al., 2002) Firstly, they have used a few words preceded by the letter waw "و "meaning "and" in 17 words together with 11 duplicates. In several words in the Arabic language, this letter comes in a different format and can come before the entire words in the language with no exceptions. One of the most appropriate methods to do this is to eliminate it through the use of an applicable and effective stemming algorithm. Secondly, they have deleted enormous single letters with the waw, specifically "ba' "ب ,"heh "ه" ."hamza "ٔ", alef "أ ,ا." Because of the nature of the written Arabic language, the aforementioned letters can come individually, but they are still considered a part of the word, so deleting them can change the meaning of the word or can make it meaningless, e.g. the word of "كتاب "that has several meanings such as writers, book, or a place of learning includes the letter ba' as a single separate letter, therefore removing it would make the word meaningless. Thirdly, a few words found are not considered stop-words although they have appeared several times in the corpus' statistics' analysis such as Casablanca " الدار البيضاء","United "المتحدة ,"States "الولايات ,"etc. Additionally, it is considered a more dependent domain list, thus it is unlikely appropriate for other collections. (Kabi et al., 2015) has categorized a group of Arabic hadiths into the so-called "Sahih AL-Bukhari", which is an 8-chapter book. It has been done by calculating the frequency of term. (Azmi et al., 2019) removed stop words using an algorithm based on the so–called "deterministic finite machine. The author recommended doing research work in the future on the impact of the steps of pre-processing such as stemming and stop-word generated from words of Hadiths. (Harrag et al., 2010) recommended deleting stop-words with high and low frequency words. As for Jbara and Khitam., 2010), he has manually helped in building a list of stop-words that includes Arabic prepositions, pronouns, names of people such as Prophet Mohammad's companions as well as places said in the corpus of hadiths. Likewise, (Harrag et al., 2014) conducted several operations that includes the elimination of stop-words required to delete the meaningless like definite articles and pronouns from the hadith Matn's text.

Other very important challenges of sentiment analysis for unstructured languages are faced in the normalization and stemming steps. Several research studies attempt to bring solutions to these problems. Indeed, (Abuata et al., 2015) developed an algorithm to remove the suffixes and prefixes of dialect words and also extract the stem of these words used in Arabian Gulf countries (Kuwait, Bahrain, Qatar, UAE, Saudi, Eastern Area, and South of Iraq). (Albogamy et al., 2016) proposed a new stemmer, for Arabic tweets, that does not rely on any root dictionary. They followed two phases: phase 1 is dedicated to producing a list of all possible stems by using the grammar, and phase 2 is selecting the shortest stem as the correct stem. They compared their stemmer with three Arabic stemmers and results showed that this one has the best accuracy. (Shoukry et al., 2012) tested the effect of pre-processing (normalization, stemming, and stop words removal) on the performance of an Arabic sentiment analysis system using Arabic and Egyptian tweets from Twitter.

(Al-Kabi et al 2015) compared two well-known Arabic stemmers and introduced a new light and heavy Arabic stemmer. (Al-Harbi et al., 2015) tested the effect of pre-processing on Saudi dialect sentiment analysis using Rapidminer. (Kanan et al., 2016) presented an algorithm containing a new set of rules for the Gulf dialects analysis. It concerns Kuwait, Bahrain, Qatar, the United Arab Emirates and parts of Saudi Arabia and some parts of southern Iraq. This new algorithm is able to handle all these Arabic dialects by defining new rules and their fusion with those currently used. It is also able to treat all non-Arabic words used in Arabic dialects. (Mulki et al., 2018) tested the impact of pre-processing techniques on sentiment analysis using three Tunisian datasets of different sizes and multiple domains. The results emphasize the positive impact of pre-processing phase in stemming, emoji recognition and negation detection tasks. (Alayba et al., 2018) studied the effect of pre-processing the text on the sentiment classification performance. Six methods of pre-processing were applied to the text: removing URLs, removing numbers, removing stop words, normalising repeated letters, normalising acronyms to their original, and normalising negative mentions. These methods were applied on five datasets and they evaluated using four classifiers. The study indicated that removing numbers, stop words, and URLs reduce the noise in the datasets. However, normalising negative words and acronyms improve the classification performance. The author of the paper applied the sentiment classification using three classes only, which are "positive", "neutral", and "negative".

4. COMPARATIVE STUDY

4.1 Criteria

In this section we developed a comparative study of the most important works conducted on Arabic dialect sentiment analysis. The comparison criteria adopted are:

- Year of the papers
- Language: before Pre-processing and after Pre-processing
- Dataset: Dataset size and Source.
- Data cleaning that deal with the noisy data.
- Normalization: which allow generating consistent word forms (normalizing repeated letters and Replace slangs, abbreviation, Emoticon...)
- Stop words: major steps conducted to remove words that are used for structuring text but do not contribute in any way to its content. Some of these words are: a, are, the, was…
- Stemming: heuristic process for deleting word affixes and leaving them in an invariant canonical form or "stem"

Validation: a model validation parameter %.

4.2 Comparison

Several studies treated the Arab dialects using different ways and mythologies. The table below gives a comparative study of these works to make a comparison based on several criteria and draw conclusions.

Ref	Language		Dataset		Data cleaning	Normalization	Stop Words	Stemming	Validation
	Before	After	Source	Size					
(Walaa et al., 2015)	Egyptian Dialect	Egyptian Dialect	FB TW	3000 comments 7000 tweet	DU, DD	__	Manual (Use Egyptian Dialect Stop word List)	__	88%
(El-Naggar et al., 2015)	Egyptian dialect	Egyptian dialect	FB	1350 comments	DRC	Replacement of Tatweel	Arabic stop words (not standard)	Light stemming for Arabic words	82.4%
(Mulki et al., 2018)	Tunisian Dialect	Tunisian Dialect		5,521 tweets from TEC 9,976 comments from TSAC 800 tweets	DRC	Replacement of emoticons	Arabic stop words (not standard) list	Khoja stemming	81.9%
(Najadat et al., 2018)	Jordanian dialect	Jordanian dialect	TW	22550 tweets	DU, DP, DD	Normalization of Hamzaa & Alef & ya	Used Arabic list Stop words	_	78.4%
(EL Abdouli et al., 2017)	Moroccan dialect and Berber Tamazight	Standard Arabic	TW	700 tweets	DU, DRC	Normalization of Hamzaa & Alef & ya	Arabic stop words (not standard) list	Stemming for Arabic words	69%
(Dahbi et al., 2018)	Moroccan dialect	Standard Arabic	NP	2000 reviews	DU, DP, DD	Normalization of Hamzaa & Alef & ya	Used Arabic list Stop words	Light stemming	83.91%
(Tachicart et al., 2018)	Moroccan dialect	Moroccan dialect	TW	6 750 tweets	DRC	Replacement of emoticons		Khoja stemming	92%
(Al-Kabi et al 2015)	Egyptian dialect	Egyptian dialect	TW	151 500 tweets	DU, DP, DD	Normalization of Hamzaa & Alef & ya	Arabic stop words (not standard)	__	93.56%
(Bettiche etal., 2018)	Algerian dialect	Algerian dialect	FB	2 000	_	-Similarity Regrouping	Arabic stop words (not standard) list	Phonetic Regrouping for dialect stemming	-
(Abdelhameed et al., 2019)	Sudan dialect	Sudan dialect	TW	11 450	DU ; DP ;	-Strength of words by calculate the repeated letters used	Used Arabic list Stop words	__	78%
(Abuata et al., 2015)	Gulf Dialect	Gulf Dialect	TW	20 345	DD; DU; DRC	extract the stem of dialect words used in Arabian Gulf countries	__	Arabic stemmer	-
(Salamah et al., 2018)	Kuwaiti Dialect	Kuwaiti Dialect	TW	340,000 tweets	DU, DP	-Normalization of Hamzaa & Alef & ya	Arabic stop words (not standard)	_	76%
(Adouane et al., 2016)	Gulf Dialect	Gulf Dialect	TW	20 345	DD, DU; DRC	extract the stem of dialect words used in Arabian Gulf	__	Arabic stemmer	-

						countries			
(Harrag et al., 2018)	Arabic dialect	MSA	__	1250	_	_	_	proposed a new stemmer	_
(Jbara et al., 2018)	Egyptian dialect	Egyptian dialect	TW	20,000 tweets		Tashkeel Tatweel Hamza Alef, lamalef, yeh, heh	Find a List of Egyptian Dialect Stop Words	light stemming using dialect words of prefixes and suffixes	78.8%
(Al ayyoub et al., 2018)	Arabic dialect	MSA	_	6081	_	_	_	introduced a new light and heavy Arabic stemmer using C#.NET languag	75.03%
(Larkey et al., 2018)	Saudi dialect	Saudi dialect	TW	5,500 tweets	DP,DN	Remove diacritics Replace ا with أ إ آ Replace ه with ة Replace ي with ى	Remove definite article (ال) Remove inseparable conjunction	Remove suffixes Remove prefixes	69%
(Kanan et al., 2016)	Gulf dialect	Gulf dialects	_	_	_	_	_	presented an algorithm containing a new set of rules for the Gulf dialects	_
(Mdhaffar et al., 2016)	Tunisian Dialect	Tunisian Dialect	FB	5 382	DU DH DS DP	replaced emoji and negation detection		Stemming algorithms on Tunisian	83%
(Abuelenin et al., 2017)	Egyptian Dialect Standard Arabic	Standard Arabic	TW	126959 tweets	RN, RU	Normalization of Hamzaa & Alef & ya Correcting misspellings	MSA stop words		95%
(Zaara et al., 2017)	Moroccan dialect	Moroccan dialect	TW	34 576 tweets	_	Strength of words by calculate the repeated letters	Moroccan Dialect - 200 Stop words		55.05 %
(Ismail et al., 2018)	Sudan dialect	Sudan dialect	TW	1200 tweets		__	built a stop-word list for Sudan dialect (626 words)	Stemming for Arabic words	-
(Oussous et al., 2018)	Moroccan dialect in Latin letters	Arabic letters Standard Arabic	FB	25 475 comments	DU; DP; DD, DRC	-Replacement of Emoticons -Normalization of Hamzaa & Alef & ya -Replacement of Acronyms & Abbreviation	Arabic stop words (not standard)	Arabic stemming	89.5%

(Oussous et al., 2020)	Moroccan dialect	Standard Arabic	NP	2000 reviews	DU; DP; DD,	Normalization of Hamzaa & Alef & ya	Used Arabic list Stop words	Khoja stemming	83.91%
(Atoum et al., 2019)	Jordanian dialect	Jordanian dialect	TW	22550 tweets	DU; DP;DD,	—	Arabic stop words (not standard)	—	78.4%
(Baly et al., 2017)	Egyptian Dialect	Standard Arabic	TW	700 tweets	DU, DP, DRC	Normalization of Hamzaa & Alef & ya	Used Arabic list Stop words	—	69%
(Duwiari et al., 2017)	Egyptian dialect	Egyptian dialect	News websites comments	1350 comments	DRC	Normalization of Hamzaa & Alef & ya	Used Arabic list Stop words	Light stemmer	83.07%
(Nahar et al., 2020)	Jordanian dialect	MSA	FB TW	2591tweets/ comments		—	Filter Stopwords (Arabic)	Arabic stemmer	
(Tartir et al., 2017)	Jordanian Dialect	Jordanian Dialect	TW	1000 tweets	DU, DSS	-Replacement of Emoticons -Normalization of Hamzaa & Alef & ya	Arabic stop words (not standard)	Stemming For Arabic words	82.1%

Table 1 : Summary of research in dialect Arabic

FB:Facebook; TW:Twitter; NP : Newspaper
DU: Deleted URLS DP: Deleted Punctuation DE: Deleted Elongation DD: Deleted Diacritics
DSS: Special symbols DRC: Repeated characters
DN: Deleted number

5. STATISTICAL ANALYSIS AND DISCUSSION

Table 1 shows that most studies have realized the pre-processing steps, but in their own way. The goal is to reduce noise and improve the data quality for more accurate results.

30.8% of studies have chosen to translate dialects into structured language in order to benefit from the wealth of work and studies carried out on these structured languages and the ease of processing them.

The **data-cleaning** step remains roughly the same for all works, used to delete non-significant data such as URLs (52.4%), as well as punctuation (43.6%), hashtags (30.5%), special characters (51.3%) …

The **stop words** elimination remains the most important step that requires a lot of effort when it concerns a dialect, unlike dealing with structured languages. Only 26.6% were able to perform automatic stop word deletion for the dialect by building a dictionary based on rules. These rules are often based on the repetition frequency of the word and its length.

After the noise is removed, the **normalization** step takes place. For Arabic dialects, the step that is often repeated is the normalization of Hamzaa & Alef & ya 7 (47.6%). On the other hand, the replacement of emoticons was only carried out by 19.04% of the works despite the subjective information which is contained in these emoticons and can help in the analysis of the feelings.

The multiple possibilities of writing a single word in a dialect make normalization more complicated. Therefore, most adopt the stemming for the structured part of comments and ignore the unstructured part. Only 33.3% provided an effort to develop rules in order to define the roots of words written in the Arabic dialects, using many techniques such as the phonetic regrouping and the similarity regrouping, which is the case for the analytical works done on the dialect of some Gulf countries and Algeria.

The detailed steps may differ from one analytical work to another, but the main parts of pre-processing remain the same. Sources of information are sometimes lost by deleting, for example, emoticons, hashtags and repeated characters, even if they represent information on the strength of word and the subjectivity of opinion.

The difficulty in the treatment of an Arabic dialect lies in the construction of the stop word dictionary automatically and the normalization by searching for the roots of the words, due to the richness of these dialects and the multiple possibilities of writing of the words in comments, whether in terms of language, word format or even Tatweel and repeated characters used sometimes to express feelings.

The lack of studies in this part also remains an important constraint, a challenge that is a driving force for effort to be made in order to further facilitate the pre-processing of these Arabic dialects.

6. CONCLUSION AND FUTURE WORKS

Sentiment analysis plays an essential role in decision-making in different fields such as politics, digital marketing (product and service evaluation), and for studying social phenomena. Because of its high value in practical applications, there has been an explosive growth in research in academia and applications in the industry.

However, there is a remarkable lack of pre-processing on unstructured languages such as the Arabic dialect "Darija as an example" even though these dialects represent a rich source of information given that they are the most used by the population on non-professional social networks.

This lack may be due to the difficulty of processing these languages, especially for stop word detection and stemming. This situation pushes us to take up the challenge and try to fill this gap in order to exploit a wealth that has not been exhausted.

Our future goals are first to benefit from the stop word detection techniques proposed in earlier works especially the automatic and semi automatic ones in order to develop stop word process detection for Moroccan Darija. Secondly, we propose a data cleaning process which takes into consideration the richness in feeling contained in certain content such as hashtags, emoticons ...and also to analyse in more detail the various works developed to realize stemming steps.

REFERENCES

Abdelhameed, H. J., & Hernández, S. M. (2019). Sentiment Analysis of Arabic Tweets in Sudanese Dialect. International Journal of New Technology and Research, 5(6), 17-22.

ABUATA, Belal et AL-OMARI, Asma. A rule-based stemmer for Arabic Gulf dialect. Journal of King Saud University-Computer and Information Sciences, 2015, vol. 27, no 2, p. 104-112.

Adouane, W., & Johansson, R. (2016, May). Gulf Arabic linguistic resource building for sentiment analysis. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 2710-2715).

Alajmi A., Saad E. M. and Darwish R. R., "Toward an ARABIC StopWords List Generation", International Journal of Computer Applications, Volume 46-No. 8 May 2012.

Alayba, A. M., Palade, V., England, M., & Iqbal, R. (2018, August). A combined CNN and LSTM model for arabic sentiment analysis. In International Cross-Domain Conference for Machine Learning and Knowledge Extraction (pp. 179-191). Springer, Cham.

AL-AYYOUB, Mahmoud, KHAMAISEH, Abed Allah, JARARWEH, Yaser, et al. A comprehensive survey of arabic sentiment analysis. Information processing & management, 2019, vol. 56, no 2, p. 320-342.

ALBOGAMY, Fahad et RAMSAY, Allan. Unsupervised stemmer for Arabic tweets. In : Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT). 2016. p. 78-84.

Al-Harbi, W. A., & Emam, A. (2015). Effect of Saudi dialect preprocessing on Arabic sentiment analysis. International Journal of Advanced Computer Technology, 4(6), 91-99.

Al-Kabi, M. N., Kazakzeh, S. A., Ata, B. M. A., Al-Rababah, S. A., & Alsmadi, I. M. (2015). A novel root based Arabic stemmer. Journal of King Saud University-Computer and Information Sciences, 27(2), 94-103.

AL-KABI, Mohammed N., WAHSHEH, Heider A., ALSMADI, Izzat M., et al. Extended topical classification of hadith Arabic text. Int. J. Islam. Appl. Comput. Sci. Technol, 2015, vol. 3, no 3, p. 13-23.

ATOUM, Jalal Omer et NOUMAN, Mais. Sentiment analysis of Arabic jordanian dialect tweets. Int. J. Adv. Comput. Sci. Appl., 2019, vol. 10, no 2, p. 256-262.

Azmi, A. M., Al-Qabbany, A. O., & Hussain, A. (2019). Computational and natural language processing based studies of hadith literature: a survey. Artificial Intelligence Review, 52(2), 1369-1414.

BALY, Ramy, EL-KHOURY, Georges, MOUKALLED, Rawan, et al. Comparative evaluation of sentiment analysis methods across Arabic dialects. Procedia Computer Science, 2017, vol. 117, p. 266-273.

Bettiche, M., Mouffok, M. Z., & Zakaria, C. (2018, June). Opinion Mining in Social Networks for Algerian Dialect. In International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (pp. 629-641). Springer, Cham.

CHEN, Aitao et GEY, Fredric. Translation term weighting and combining translation resources in cross-language retrieval. In : TREC. 2001. p. 2001.

Dahbi, M., Saadane, R., & Mbarki, S. (2019, October). Citizen Sentiment Analysis in Social Media Moroccan Dialect as Case Study. In The Proceedings of the Third International Conference on Smart City Applications (pp. 16-29). Springer, Cham.

DUWAIRI, Rehab M. Sentiment analysis for dialectical Arabic. In : 2015 6th International Conference on Information and Communication Systems (ICICS). IEEE, 2015. p. 166-170.

El Abdouli, A., Hassouni, L., & Anoun, H. (2017). Sentiment analysis of moroccan tweets using naive bayes algorithm. International Journal of Computer Science and Information Security (IJCSIS), 15(12).

El-Naggar, N., El-Sonbaty, Y., & Abou El-Nasr, M. (2017, July). Sentiment analysis of modern standard Arabic and Egyptian dialectal Arabic tweets. In 2017 Computing Conference (pp. 880-887). IEEE.

G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. (references)

Gharatkar, S., Ingle, A., Naik, T., & Save, A. (2017, March). Review preprocessing using data cleaning and stemming technique. In 2017 international conference on innovations in information, embedded and communication systems (iciiecs) (pp. 1-4). IEEE.

HARRAG, Fouzi et AL-QAWASMAH, Eyas. Improving Arabic Text Categorization Using Neural Network with SVD. J. Digit. Inf. Manag., 2010, vol. 8, no 4, p. 233-239.

HARRAG, Fouzi. Text mining approach for knowledge extraction in Sahîh Al-Bukhari. Computers in Human Behavior, 2014, vol. 30, p. 558-566.

ISMAIL, Rua, OMER, Mawada, TABIR, Mawada, et al. Sentiment analysis for arabic dialect using supervised learning. In : 2018 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE). IEEE, 2018. p. 1-6.

JBARA, Khitam. Knowledge discovery in Al-Hadith using text classification algorithm. Journal of American Science, 2010, vol. 6, no 11, p. 409-419.

KANAN, Tarek et FOX, Edward A. Automated arabic text classification with P-S temmer, machine learning, and a tailored news article taxonomy. Journal of the Association for Information Science and Technology, 2016, vol. 67, no 11, p. 2667-2683.

KHOJA, Shereen et GARSIDE, Roger. Stemming arabic text. Lancaster, UK, Computing Department, Lancaster University, 1999.

LARKEY, Leah S. et CONNELL, Margaret E. Arabic information retrieval at UMass in TREC-10. In : TREC. 2001
LARKEY, Leah S., BALLESTEROS, Lisa, et CONNELL, Margaret E. Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis. In : Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. 2002. p. 275-282.

LIU, Kun-Lin, LI, Wu-Jun, et GUO, Minyi. Emoticon smoothed language models for twitter sentiment analysis. In : Aaai. 2012. p. 22-26.

Mdhaffar, S., Bougares, F., Esteve, Y., & Hadrich-Belguith, L. (2017, April). Sentiment analysis of tunisian dialects: Linguistic ressources and experiments.

MELVILLE, Prem, GRYC, Wojciech, et LAWRENCE, Richard D. Sentiment analysis of blogs by combining lexical knowledge with text classification. In : Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. 2009. p. 1275-1284.

MULKI, Hala, HADDAD, Hatem, BECHIKH ALI, Chedi, et al. Tunisian dialect sentiment analysis: a natural language processing-based approach. Computación y Sistemas, 2018, vol. 22, no 4.

Nahar, K. M., Jaradat, A., Atoum, M. S., & Ibrahim, F. (2020). SENTIMENT ANALYSIS AND CLASSIFICATION OF ARAB JORDANIAN FACEBOOK COMMENTS FOR JORDANIAN TELECOM COMPANIES USING LEXICON-BASED APPROACH AND MACHINE LEARNING. Jordanian Journal of Computers and Information Technology (JJCIT), 6(03).

Najadat, H., Al-Abdi, A., & Sayaheen, Y. (2018, April). Model-based sentiment analysis of customer satisfaction for the Jordanian telecommunication companies. In 2018 9th International Conference on Information and Communication Systems (ICICS) (pp. 233-237). IEEE.

Nassr, Z., Sael, N., &Benabbou, F. (2019, October). A comparative study of sentiment analysis approaches. In Proceedings of the 4th International Conference on Smart City Applications (pp. 1-8).

NASSR, Zineb, SAEL, Nawal, et BENABBOU, Faouzia. Machine Learning for Sentiment Analysis: A Survey. In : The Proceedings of the Third International Conference on Smart City Applications. Springer, Cham, 2019. p. 63-72.

OUSSOUS, Ahmed, BENJELLOUN, Fatima-Zahra, LAHCEN, Ayoub Ait, et al. ASA: A framework for Arabic sentiment analysis. Journal of Information Science, 2020, vol. 46, no 4, p. 544-559.

OUSSOUS, Ahmed, LAHCEN, Ayoub Ait, et BELFKIH, Samir. Improving sentiment analysis of Moroccan tweets using ensemble learning. In : International Conference on Big Data, Cloud and Applications. Springer, Cham, 2018. p. 91-104.

Salamah, J. B., & Elkhlifi, A. (2014, January). Microblogging opinion mining approach for kuwaiti dialect. In The International Conference on Computing Technology and Information Management (ICCTIM) (p. 388). Society of Digital Information and Wireless Communication.

SAVOY, Jacques et RASOLOFO, Yves. Report on the TREC 11 experiment: Arabic, named page and topic distillation searches. In : TREC. 2002.

Shoukry, A., & Rafea, A. (2012, November). Preprocessing Egyptian dialect tweets for sentiment mining. In The Fourth Workshop on Computational Approaches to Arabic Script-based Languages (p. 47).

SOUMEUR, Assia, MOKDADI, Mheni, GUESSOUM, Ahmed, et al. Sentiment analysis of users on social networks: overcoming the challenge of the loose usages of the Algerian Dialect. Procedia computer science, 2018, vol. 142, p. 26-37.

Tachicart, R., & Bouzoubaa, K. (2019, September). An empirical analysis of Moroccan dialectal user-generated text. In International Conference on Computational Collective Intelligence (pp. 3-12). Springer, Cham.

TAN, Chenhao, LEE, Lillian, TANG, Jie, et al. User-level sentiment analysis incorporating social networks. In : Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. 2011. p. 1397-1405.

TARTIR, Samir et ABDUL-NABI, Ibrahim. Semantic sentiment analysis in Arabic social media. Journal of King Saud University-Computer and Information Sciences, 2017, vol. 29, no 2, p. 229-233.

Walaa M, Ahmed Y and Hoda K, “Egyptian Dialect Stopword List Generation from Social Network Data”,Egyptian Journal of Language Engineering, Vol 2, No. 1, April 2015

ZARRA, Taoufiq, CHIHEB, Raddouane, MOUMEN, Rajae, et al. Topic and sentiment model applied to the colloquial Arabic: a case study of Maghrebi Arabic. In : Proceedings of the 2017 international conference on smart digital environment. 2017. p. 174-181.