EasyChair Preprint

№ 4069

Preprocessing Arabic Dialect for Sentiment Mining: State of Arte

Zineb Nassr, Nawal Sael and Faouzia Benabbou

August 23, 2020

Preprocessing Arabic dialect for sentiment mining: state of arte

Zineb NASSR
Laboratory of Modelling and Information Technology
Faculty of Sciences Ben M'SIK, University Hassan II
Casablanca, Morocco
nassrzineb@gmail.com

Nawal SAEL
Laboratory of Modelling and Information Technology
Faculty of Sciences Ben M'SIK, University Hassan II
Casablanca, Morocco
saelnawal@hotmail.com

Faouzia BENABBOU
Laboratory of Modelling and Information Technology
Faculty of Sciences Ben M'SIK, University Hassan II
Casablanca, Morocco
faouzia.benabbou@univh2c.ma

***Abstract*—Sentiment Analysis, concerns with 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 the opinions of users about various products or topics discussed over 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 presents a literature review on the preprocessing, an analytical and comparative study of different researches conducted on sentiment analysis in Arabic social networks. Our comparison analyzes in more detail the preprocessing (stop words, stemming…) steps which are very important in sentiment analysis process success and are the most difficult especially in the case where the comments are written in Arabic dialect.**

Keywords— Preprocessing, Arabic Dialect, Sentiment mining, Stop-words, Stemming

INTRODUCTION

This Today social networks have become in some way one of the most popular communication tools. These 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 file opinions, share experiences or simply talk about everything about them online [1].The monitoring of social media has become an important way for analyzing and detecting trends, studying and evaluating opinions on various topics such as politics [2], services (teachings, health...), marketing [3] and business products.

People can share their opinions in an environment without constraint and, companies can extract useful ideas for their decision-making process [5]. To quantify what people think from textual qualitative data, a polarity classification task for detecting positive, negative, or neutral text is required [6]. Although, the large number of research available on the analysis of documents such as newspaper articles, journals, there are still several open questions to tackle on 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 are dedicated to provide a precise classification of sentiment for unstructured languages in general and more specifically for the Moroccan dialect "Darija". In our last work [7] we carried out a state of the art and a comparative study of research done in recent years on sentiment analysis .In addition to other finding, we deduced that, most of researches translate the comments in a structured language such as English to analyze them and that there are no standard resources for unstructured languages, in particular for Maroccan Darija (MDL). Specially, with the fact that, Comments in Darija can be wrote in Arabic characters, in Latin character or both of them. This makes automatic processing more difficult to achieve.

In fact, user-generated content on the Web is generally unstructured and need important preprocessing steps and analysis to extract useful knowledge [8]. These steps depend on the nature of the language (structured or unstructured) and generally different from one research to another. Their objective is 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, preprocessing of the text, sensing of sentiment, classification of sentiment [9]. Nevertheless, the preprocessing step is the most important in the analysis of feelings because the messages in the social networks are characterized by expressions colloquial, abbreviations, emoticons [10], a lengthening of words, a capital letter irregular and they are not generally conform to the canonical grammatical rules.

Preprocessing phase is faced to 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 in online social media platforms and is often replaced by some language conventions. The lengthening of words like "it'sseeeeeerious" (commonly known as expressive elongation or Word stretching) is an example of new language conventions that are today very popular in online platforms [11]. Other problems are related to additional terms such as the abbreviation expressions that are additional paralinguistic elements used in non-verbal

communication in online social networks [12]. The Hashtags **w**hich are widely used in online social networks [13] 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 [14] to reflect the role played by facial expressions in speech.

Other very important preprocessing challenge is to detect and analyze the uppercase letters since the 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 [15].

Il faut mettre un paragraphe ici qui rappelle l'objectif du travail, et la structure du papier (comment il est organisé)

PREPROCESSING BACKGROUND

A. Preprocessing task Major steps

To overcome preprocessing challenges, This phase can be carried out in several steps which depend on the nature of the language and the analysis objectives. The major steps are:

Data cleanin : parler brievement de son role et certaines particularité du domaine

Preprocessing

- Remove diacritics
- Remove definite article (ال)
- Remove inseparable conjunction (و)
- Remove suffixes and prefixes

Stopword 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 [16].

Stemming: heuristic process for deleting word affixes and leaving them in an invariant canonical form or "stem" [17]. For instance, person, person's, personify and personification become person when stemmed. The most popular English stemmer algorithm is Porter's stemmer [18].

Lemmatization: algorithmic process to bring 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 [19].

B. Arabic dialect preprocessing challenges

- Preprocessing phase is faced to several problems, which are related to the sentiment analysis context. Match result against a list of patterns. If a match is found, extract the characters in the pattern representing the root.
- .Match the extracted root against a list known "valid"‖ roots
- Replace weak letters و ا ي with و
- Replace all occurrences of Hamza ؤ ء ئ with ا
- Two letter roots are checked to see if they should contain a double character. If so, the character is added to the root.

Table 1 shows different challenges in preprocessing the Arabic dialect.

Dialect challenges	**Example**	**Preprocessing Problem**
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	In MSA ("كتاب kitAb") In MorD ("كتاب ktAb") In English (Book)	Problem with stemming , so we must include all these possible forms of writing
Bypassing or avoiding the Hamza to be pronounced as a "ya"	In MSA ("مائدة MA'ida") In MorD ("مايده MAydah") In English (Table)	Same problem in stemming
Some pronouns are slightly modified from their MSA	ntouma") for "أنتم ntouma") (« ntiya (« نت or نتي nti») instead of انت») anti »)	These are stop word modified in Moroccan dialect, so they are not covered by the stop word dictionary of the Arabic language.
For the possessive, it is common to add the word ديالي) my) instead of just the MSA suffixed pronoun.	كتابي) ktAby") (كتاب ديالي My book	Instead of having a single word which will be processed by tokenization, we will have an additional word which will be considered as a stop word and it must be taken over in the dictionary of stop words to be deleted.
Some of the interrogative particles are slightly modified.	فين) fyn"fen"fin" for "where"), شكون) shkoun"chkon" for "who"),	Another example of stop word that does not comply with their correspondents in Arabic, hence the difficulty of detecting them
The negation is introduced by means of the word ما ») mA ») and the suffix ش with the sukun on it (« sh »)	as in ممشيتش .Negation also has some other expressions such as) mA rAnysh, "I am not") or ماراناش) marAnAsh, "We are not"), etc;	Difficulty detecting negation since it will be represented in the same word with one more prefix and suffix.
a number of words that are not of Arabic origin have percolated into MorD,	such as طابلة ("TAblah") for Table from French,''كارطبلي**") ca**rTably") for "my schoolbag", from French, etc.	More words which are not part of the Arabic language, but which are of French origin, which broadens the basis for tokenization
Words contain numbers instead of letters	Mo9ati3one for Mokatiône	Characters that are expressed as numbers that must be taken care of in stop word and tokenization.

The above, especially the part related to the use of words of non-Arabic origin and Stop word, already highlights some of the Challenges of preprocessing MorD.

RELATED WORKS

Several studies have been interested in sentiment analysis and a variety of approaches has been developed, especially for English language. However, researches are more limited for other languages including Arabic. This section discusses researches in the field of sentiment analysis for Arabic dialects.

Be Bilal Abta and Asma Al-Omari [4] 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 eastern Saudi Arabia and some parts of southern Iraq. This new algorithm is able to handle all known Arabic dialects by defining new rules and the fusion of these rules with the rules currently used and also to treat all non-Arabic words used in Arabic dialects

Assia Soumeur et al [20] tackled the problem of SA of Algerian Facebookers' comments on published pages belonging to various companies. First, they studied the specificity of Algerians Dialect and the linguistic behaviour of Algerians on social media. They built a corpus of more than 25000 sentiment-annotated comments. This corpus then went through a preprocessing phase each step of which was evaluated in terms of its impact on the quality of the data using a Naïve Bayes classifier. They noticed the strong impact of this phase on the results they obtained, the performance significantly increasing after each step. Two types of deep neural networks have been implemented. The first is a (deep) MLP where the best configuration gave an accuracy of 81.6%. The second is a Convolutional Neural Network that reached an accuracy of 89.5%. These, they believe, are very encouraging results given the complexity of AlgD on social media

Huda Jamal et al [21] focused on sentiment analysis of Arabic tweets written using either Modern Standard Arabic or Sudanese dialectical Arabic. They have created their own lexicon which contain 2500 words and they have applied three different classifiers

In [23], the authors exploited the NB classifier to help classify Arabic Facebook posts informally written and in different dialects (mainly Syrian, Egyptian, Iraqi and Lebanese). Results show that better results are achieved when Naïve search is used as a binary classifier to classify the Facebook posts as being either objective or subjective. The authors extended their classification to include spam and dual sentiments (posts including two different sentiments).

The authors of [24] proposed a colloquial Non-Standard Arabic-Modern Standard Arabic-Sentiment Analysis Tool (CSA-MSA-SAT) that is used to determine the polarities of different colloquial Arabic and MSA reviews and comments using 18 manually built polarity lexicons (9 domain specific positive polarity lexicons, 9 domain specific negative polarity lexicons, and 2 general purpose lexicons). The domains covered are: technology, books, education, movies, places, politics, products, and society. These lexicons consist of 1,080 Arabic reviews from around 70 different social media and news websites. The algorithm used is based on the sentiment term frequencies to identify the subjectivity and the polarity of different Arabic reviews and comments. Experimental results showed a 90% accuracy using K-Nearest Neighbor (KNN) classifier.

The authors of [25] studied three lexicon construction techniques, one manual technique and two automatic. In addition, an Arabic SA tool is designed. Experiments showed that using the lexicon of 16,800 word created by integrating the three different lexicon construction techniques detailed in their study is the most useful. The accuracy of implementing the Arabic SA tool was 74.6%

The authors of [26] proposed an approach that automatically classifies different opinions written in MSA or/and colloquial Arabic into a predefined set of categories. The authors used the same 18 manually built lexicons used in [7]. Three algorithms were proposed: a subjectivity algorithm to categorize Arabic reviews as facts or opinions, a polarity algorithm to determine the polarity (positive, negative, neutral or undetermined) of evaluated Arabic reviews, and a strength/intensity algorithm used to determine the strength/intensity (strong positive, strong negative, weak positive, weak negative, neutral or an undetermined review) of evaluated Arabic reviews.

Other researches were conducted on Arabic tweets. The authors of [27] developed a framework that determines the polarity of Arabic Tweets. A primary contribution of the authors is that they provide lexicons making it possible to handle Jordanian dialects, Arabizi and emoticons. The authors applied three classifiers NB, SVM and KNN on 25000+ tweets and experimental results showed that the highest accuracy is 76.78% achieved by the NB classifier without using the stop words filter and stemming with folds of 5.

The authors of [28] proposed a lexicon-based computationally light method for sentiment analysis of Arabic tweets on mobile devices. The proposed method classifies the tweet into positive, negative, objective or neutral using decision trees as the classification model. Experiments were conducted using a corpus of 2300 manually annotated Arabic Tweets. A SC accuracy of 67.3% was obtained.

In [29], the researchers combined different approaches to design and implement a subjective SA system. The authors also presented an improvement to the hybrid approach proposed in [30], and this is done using the merged lexicon, along with a modified semantic orientation algorithm and then applying the feature selection using the Information Gain measure. Results showed an 84% polarity classification accuracy and an 81% subjectivity classification accuracy.

In [30]; the two components are a lexicon and an SA tool. The authors started by building an Arabic corpus of 4000 textual comments, collected from twitter and Yahoo!-Maktoob. An Arabic lexicon is then built from a seed of 300 words, positive words are given the weight +1 and negative words are given the weight -1. Next, the lexicon-based tool is created. The accuracy of this tool reaches 70.05% without light stemming and only on twitter comments, however the accuracy on Yahoo!-Maktoob without light stemming is 63.75%.

The author of [31] proposed a SA and Negation Resolving system for Arabic Text Entailment (SANATE).

For experimentation, the dataset for Arabic textual entailment (ArbTEDS) consisted of 618 text-hypothesis pairs. Each texthypothesis pair is entered in the ATE system and SANATE system. The accuracy of ATE is 0.617 and the accuracy of SANATE is 0.693. These results show that resolving the negation and classifying text polarity, increases the performance of detecting the entailment relation and nonentailment relation.

ABUATA et al [32] extracted the stem of dialect words used in Arabian Gulf countries (Kuwait, Bahrain, Qatar, UAE, Saudi, Eastern Area, and South of Iraq.

1. A COMPARATIVE STUDY

2.1. Criteria

In our comparative study, we regroup and synthesize the researches done in the last few years. The comparison criteria adopted are :

- **Year:**
- **Language**: before Preprocessing and after Preprocessing
- **Dataset**: Size and Source of dataset.
- **Data cleaning** that deal with the noisy data;
- **Normalization**: which allow to generate consistent word forms (Stemming, Lemmatisation, Normalizing repeated letters and Replaceslangs
- **Stop words:** 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…
- **Stemming :** heuristic process for deleting word affixes and leaving them in an invariant canonical form or "stem"
- **Validation**: a validation parameter of precision

Table 1 Summary of research in dialect Arabic

Ref	Year	Language		Dataset		Data cleaning	Normalization	Stop Words	Stemming	Validation
		Before	After	Source	Size					
[33]	2019	French letters	Arabic letters	FB	25 475 comments	DURL; DP;DD, DRC	- Remplacement Emoticones -Normalisation of Hamzaa & Alef & ya Remplacement Acronyms Abbreviation	Automatic	NO	89.5%
[34]	2014	Egyptian dialect	Egyptian dialect	FB	1350 comments	DRC	Remplacement of Tatweel	Manual	YES	82.4%
[35]	2017	Jordanian Dialect	Jordanian Dialect	TW	1000 tweets	DU,DSS	- Remplacement Emoticones -Normalisation of Hamzaa & Alef & ya	Manual	YES	82.1%
[36]	2018	Kuwaiti Dialect	Kuwati Dialect	TW	340,000 tweets	DU,DP	Normalisation of Hamzaa & Alef & ya	Automatic	NO	76%
[37]	2015	Jordanian dialect	Jordanian Arabic dialect	TW	1000 tweets	DU; DP;DD,		Manual	NO	75%
[38]	2017	Tunisian Dialect	Tunisian Dialect	FB TW	5,521 tweets from TEC 9,976 comments from TSAC 800 tweets	DRC	Remplacement of emoticons	Automatic	YES	81.9%

[39]	2018	Jordanian dialect	Jordanian dialect	TW	22550 tweets	DU; DP;DD,		Manual	NO	78.4%
[40]	2019	moroccan dialect and Berber Tamazight	Standard Arabic	TW	700 tweets	DRC	Normalisation of Hamzaa & Alef & ya	Manual	NO	69%
[41]	2018	Moroccan dialect	Standard Arabic	NP	2000 reviews	DU; DP;DD,	Normalisation of Hamzaa & Alef & ya	Automatic	NO	83.91%
[42]	2018	Moroccan dialect	Moroccan dialect	TW	6750 tweets	DRC	Remplacement of emoticons	Manual	YES	92%
[43]	2018	Egyptian dialect	Egyptian dialect	TW	151,500tweets	DU; DP;DD,	Normalisation of Hamzaa & Alef & ya	Manual	NO	93.56%
[45]	2017	Jordanian and Saudi Arabia dialect	Jordanian and Saudi Arabia dialect	NP	28,576 reviews	DRC	Normalisation of Hamzaa & Alef & ya	Automatic	NO	95%

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

C. Analysis study

Most studies have followed the steps of pre-treatment, but in their own way. the goal is to reduce noise for more accurate results. Pretreatment in all the work begins with a step of cleaning data, conducting a removal of URLs, hashtags, repeated characters, emoticons and special characters ... with differences from one job to another of the elements included in this list, first cleaning layer is carried out on the relevant datasets. Subsequently, among the most important elements, the removal of words stops is done manually in some work or automatically on others by building a stop word dictionary based on rules. This step represents a great challenge and a remarkable effort when it comes to work performed on a dialect, unlike the works that deal with structured languages. Then, the normalization step will take place, with a broad base of opportunities that can be used to write the words in a dialect, the task becomes complicated, that's why most of the work is a right stemming from the structured part and ignore the unstructured part, if other they make a further effort to develop rules to find the roots of the words of the informal dialect as the case of the analytical work that has been done on the dialect of the Gulf countries and that of Egypt.

These steps differ from one work to another and will impact the final outcome, sources of information in the Data cleaning or level of standardization, can be lost as emoticons, hashtags that can give an idea about subjectivity of opinion, or even the repetition of characters that can also provide information on the strength of this opinion, however, the overall goal is to remove noise and see

a clean dataset on which a classification can be applied correct.

CONLCUSION

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 for practical applications, there has been an explosive growth in research in academia and applications in the industry.

However, there is a remarkable lack of treatment on unstructured languages such as the Arabic dialect "Darija as an example", yet These dialects represent a rich source of information especially 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 in terms of pretreatment, something that pushes us to take up the challenge and try to fill this gap in order to exploit a little used wealth.

In a future enhancement, we can use other criteria more for annotation on which we do not focus in this works. Analysis

of hashtags and emoticons in addition to the texts will be one more tool for annotating comments and repeating words altogether and not just the characters that we may be also useful for assigning depth

References

[1] 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)*

[2] 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).

[3] Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. Tourism Management, 58, 51-65.

[4] Mavridis, N. (2015). A review of verbal and non-verbal human–robot interactive communication. Robotics and Autonomous Systems, 63, 22-35.

[5] [4]Nazir, F., Ghazanfar, M. A., Maqsood, M., Aadil, F., Rho, S., &Mehmood, I. (2019). Social media signal detection using tweets volume, hashtag, and sentiment analysis. Multimedia Tools and Applications, 78(3), 3553-3586.

[6] [5]Wang, H., &Castanon, J. A. (2015, October). Sentiment expression via emoticons on social media. In 2015 ieee international conference on big data (big data) (pp. 2404-2408). IEEE.

[7] [6]Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

[8] [7]Nassr, Z., Sael, N., &Benabbou, F. (2019, October). Machine Learning for Sentiment Analysis: A Survey. In The Proceedings of the Third International Conference on Smart City Applications (pp. 63-72). Springer, Cham.

[9] 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.

[10] ABDELHAMEED, Huda Jamal et MUÑOZ-HERNÁNDEZ, Susana. Sentiment Analysis of Arabic Tweets in Sudanese Dialect.

[11] SHOUKRY, Amira et RAFEA, Ahmed. Preprocessing Egyptian dialect tweets for sentiment mining. In : *The Fourth Workshop on Computational Approaches to Arabic Script-based Languages*. 2012. p. 47.

[12] 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.

[13] . Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[14] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[15] FUNG, Pascale. Extracting key terms from Chinese and Japanese texts. Computer Processing of Oriental Languages, 1998, vol. 12, no 1, p. 99-121.

[16] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[17] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[18] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[19] Khan K, Baharudin B, Khan A and Malik F. Mining opinion from text documents: A survey. In: Proceedings of the 3rd IEEE

[20] international conference on digital ecosystems and technologies, Istanbul, 2009, pp. 217–222. [8] Web. Top 30 Languages of the World, http://www.vistawide.com/languages/top_30_languages.htm (accessed 3 September

[21] 2013).

[22] [9] Saleh MR, Martı´n-Valdivia MT, Uren˜a-Lo´pez LA and Perea-Ortega JM. OCA Corpus English Version, http://sinai.ujaen.es/

[23] wiki/index.php/OCA_Corpus_(English_version) (accessed 3 September 2013).

[11] Pang B, Lee L and Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings

[24] of the conference on empirical methods in natural language processing (EMNLP ’02), Morristown, 2002, pp. 79–86.

[25] [12] Bickerstaffe A and Zukerman I. A hierarchical classifier applied to multi-way sentiment detection. In: Proceedings of the 23rd

[26] international conference on computational linguistics, Beijing, 2005, pp. 62–70.

[27] [13] Paltoglou G and Thelwall M. A study of information retrieval weighting schemes for sentiment analysis. In: Proceedings of the

[28] 48th annual meeting of the association for computational linguistics, Uppsala, 2010, pp. 1386–1395.

[29] [14] Hall M. Correlation-based feature selection for machine learning. PhD thesis, University of Waikato, 1999.

[30] [15] Esuli A and Sebastiani F. Determining term subjectivity and term orientation for opinion mining. In: Proceedings of the 11th

[31] conference of the European Chapter of the Association for Computational Linguistics, Trento, 2006, pp. 193– 200.

[32] HEDAR, Abdel Rahman et DOSS, M. Mining social networks arabic slang comments. In : IEEE Symposium on Computational Intelligence and Data Mining (CIDM). 2013.

[33] 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.

[34] SALAMAH, Jana Ben et ELKHLIFI, Aymen. Microblogging opinion mining approach for kuwaiti dialect. In : *The International Conference on Computing Technology and Information Management (ICCTIM)*. Society of Digital Information and Wireless Communication, 2014. p. 388.

[35] 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.

[36] 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.

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

[38] EL ABDOULI, Abdeljalil, HASSOUNI, Larbi, et ANOUN, Houda. Sentiment Analysis of Moroccan Tweets using Naive Bayes Algorithm. International Journal of Computer Science and Information Security (IJCSIS), 2017, vol. 15, no 12.

[39] 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.

[40] DAHBI, Monir, SAADANE, Rachid, et MBARKI, Samir. Social media sentiment monitoring in smart cities: an application to Moroccan dialects. In : Proceedings of the 4th International Conference on Smart City Applications. 2019. p. 1-6.

[41] DONIAGAMAL, Marco Alfonse, EL-HORBATY, El-Sayed M., et SALEM, Abdel-BadeehM. Opinion mining for Arabic dialects on twitter. *Egyptian Computer Science Journal*, 2018, vol. 42, no 4.

[42] DONIAGAMAL, Marco Alfonse, EL-HORBATY, El-Sayed M., et SALEM, Abdel-BadeehM. Opinion mining for Arabic dialects on twitter. Egyptian Computer Science Journal, 2018, vol. 42, no 4.

[43] 30 BETTICHE, Mehdi, MOUFFOK, Moncef Zakaria, et ZAKARIA, Chahnez. Opinion Mining in Social Networks for Algerian Dialect. In : International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. Springer, Cham, 2018. p. 629-641.

[44] 31 SAFEEK, Ilham et KALIDEEN, Muhammad Rifthy. Preprocessing on Facebook data for sentiment analysis. 2017.

[45] 32 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.

[46] 33 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.

[47] [16] Sarvabhotla K, Pingali P and Varma V. Sentiment Classification: A Lexical Similarity Based Approach for Extracting

[48] Subjectivity in Documents. Information Retrieval 2011; 14(3): 337–353.

[49] [17] Denecke K. Are SentiWordNet scores suited for multi-domain sentiment classification? In: Proceedings of the fourth international conference on digital information management (ICDIM), Ann Arbor, MI, 2009, pp. 33–38.

[50] [18] Ohana B and Tierney B. Sentiment classification of reviews using SentiWordNet. In: Proceedings of the 9th information technology and telecommunication conference (IT & T), Dublin, 2009.

[51] [19] Esuli A and Sebastiani F. SentiWordNet: A publicly available lexical resource for opinion mining. In: Proceedings of the 5th

[52] international conference on language resources and evaluation (LREC), Genova, 2006, pp. 417–422.

[53] [20] Denecke K. Using SentiWordNet for multilingual sentiment analysis. In: Proceedings of IEEE 24th international conference on

[54] data engineering workshop (ICDEW), Hannover, 2007, pp. 507–512.

[55] [21] Kosinov S. Evaluation of N-grams conflation approach in text-based information retrieval. In: Proceedings of international

[56] workshop on information retrieval, Edmonton, Alberta, 2001, pp. 136–142.

[57] [22] El-Halees A. Arabic opinion mining using combined classification approach. In: Proceedings of the international Arab conference on information technology (ACIT), Riyadh, 2011.

[58] [23] Thelwall M, Buckley K, Paltoglou G, Cai D and Kappas A. Sentiment strength detection in short informal text. Journal of the

[59] American Society for Information Science and Technology 2010; 61(12): 2544–2558.

[60] [24] Ahmad K and Almas Y. Visualizing sentiments in financial texts. In: Proceedings of the ninth international conference on information visualization, Washington, DC, 2005, pp. 363–368.

[61] [25] Ahmad K, Cheng D and Almas Y. Multi-lingual sentiment analysis of financial news streams. In: Proceedings of the 1st international workshop on grid technology for financial modeling and simulation, Palermo, 2006.

[62] [26] Abbasi A, Chen H and Salem A. Sentiment analysis in multiple languages: Feature selection for opinion classification in web

[63] forums. ACM Transactions on Information Systems 2008; 26(3): 1–34.

[64] [27] Elhawary M and Elfeky M. Mining Arabic business reviews. In: Proceedings of the IEEE international conference on data mining, Mountain View, CA, 2010, pp. 1108–1113.

[65] [28] Saleh MR, Martı´n-Valdivia MT, Uren˜a-Lo´pez LA and Perea-Ortega JM OCA: Opinion corpus for Arabic. Journal of the

[66] American Society for Information Science and Technology 2011; 62(10): 2045–2054.

[67] [29] Rahmoun A and Elberrichi Z. Experimenting N-grams in text categorization. The International Arab Journal of Information

[68] Technology 2007; 4(4): 377–385.

[30] Wu X et al. Top 10 Algorithms in data mining. Knowledge and Information Systems 2008; 14(1): 1– 37.