Multilingual sentiments analysis to improve the quality of services provided by Algerian telephone operator

Klouche Badia[1]and Benslimane Sidi Mohammed[1]

[1]Ecole Supérieure en Informatique, LabRi Laboratory, Sidi Bel Abbès, Algeria

b.klouche@esi-sba.dz
s.benslimane@esi-sba.dz

Abstract.
In the world of BI, the importance of facts is undeniable. Sentiment Analysis (SA) is a research area at the crossroads of many areas, such as data mining, natural language processing, and machine learning. This is the automatic extraction of opinions expressed in a given text. Due to its many applications, numerous studies have been conducted in the field of Opinion Manning. Most approaches in this area are focused on English because of the lack of trusted resources in other languages such as the Arabic language and its great diversity dialects, especially in texts in the Algerian Arabic dialect. As other companies, Algerian telephone operators, attach great importance to the opinion of their customers. Nonetheless, customers generally use Algerian Arabic dialect to answer Short Message Services (SMS) questionnaires.
In this work, we propose multilingual sentiments analysis approach based on the feedback of the customers of the Telephone operator Ooredoo, written in Modern Standard Arabic, Arabizi or Algerian dialect. The proposed approach permit to the operator to improve the quality of its services in order to conquer new customers and expand its usual clientele.

Keywords: sentiment analysis, machine learning, text classification, opinion mining and sentiment, dialectal Arabic.

1 Introduction

Nowadays, with the advent of web 2.0, the digital world is recording billions of users of social media sites and applications [1].In addition, the analysis of sentiments (SA) becomes a field of study very open to research, whose objective is to analyze, from texts shared on social networks, opinions, feelings, attitudes and emotions on different topics [2]. Indeed, it has always been very important to know the opinions of others, on various issues, such as: products, services and organizations. These opinions are all the more relevant because they come from people who have experienced the product or service in question. Report that nearly 87% of online reviewers of restaurants, hotels and other services admit that they have had some influence on their consumption. In the field of telephony, telephone operators use a combination of intuition, experience and a certain level of analysis to make strategic and tactical decisions.

As a result, the majority of companies in the Sector frequently use effective client voice collection and listening programs such as: (1) the direct surveys carried out by the communication boxes, (2) focus groups, (3) questionnaires by Short Message Services (SMS), (4) telephone surveys,(5) collections by internet or post.
Often, these traditional and costly surveys provide outdated information.
In the literature, there are many publications related to opinion polls in different disciplines, most of whose works are adapted to the English language. Arabic is the official language of more than 20 countries, its inflection system is very rich and it is considered one of the most adapted languages in terms of texture. By exploiting the research related to the analysis of the feelings of the Arabic language, several authors have determined that its progress was very slow and that they clearly lacked tools and resources for analysis of the Arabic language. However, this rich language becomes very interesting for many researchers working in the field of text extraction and information retrieval [3]. Several studies have been conducted in this context, where different corpora, resources and tools are available to test and implement applications, such as text classification. These methods can be divided into three large families, of which it should be mentioned that the first is based on machine learning (supervised methods), the second is based on the lexicon (unsupervised methods) and the third relates to hybrid approaches.
In this work, we are interested in the use of a semantic approach, combined with automatic learning for the Analysis of Feelings listed from the comments written in Arabic Dialectal Algerian, collected at the level of the Algerian Telephone Operator Ooredoo.
The rest of this paper is organized as follows: Section 2 describes the state of the art with a comparative table on the different approaches and their applications to written comments in Arabic Dialectal Algerian (ADA). Section 3 deals with the proposal of our architecture of the Analysis of Feelings, methods of selection and extraction of variables (words or sequences of words) used in the classification phase. A conclusion and perspectives of this work are presented in section 4.

2 Sentiment analysis: Background information

Sentiments analysis or opinion mining is the area of study that analyzes people's opinions, feelings, evaluations, emotions from written language. It is one of the most active areas of research in the field of natural language processing, data mining, web search, and text mining [2].
The analysis of feelings can be analyzed at three levels of granularity: document level, sentence level and aspect level. The SA at the document level is based on the assumption that it expresses a single opinion towards a single entity from the same source. The main task is therefore to determine the general orientation of the feeling of the document according to the classes that can be positive, negative or neutral [2].
At the level of the sentence, the analysis of feelings aims to identify whether it has an opinion or not and to assess the orientation of subjective sentences by feeling. This level of granularity is all the more problematic as the orientation of the words, based

on the feeling strongly depends on the context. The classification of sentiments at the level of the sentence also deals with comparative and sarcastic sentences.
The level of appearance finally, makes a finer analysis. For this one, which is more complete, it is necessary to detect the aspects of a subject and to determine by that, the feelings relative to these last ones.
The objective is to discover all the quintuple (carrier, object, appearance, feeling, time) in a given document. For example, in the sentence "The image quality of the camera is great, but it is very expensive", the analysis of the feelings at the "aspect level" must detect a positive feeling towards the aspect "image quality" as well than a negative feeling towards the "price" aspect.

3 Approaches for sentiments analysis

Several approaches have been adopted in the literature to determine the polarity of a text. These can be divided into three main families that are supervised, unsupervised and hybrid approaches.
Supervised approaches are based on machine learning algorithms, such as the Support Vector Machine (SVM), Naïve Bayes (NB).
Unsupervised approaches are lexicon-based approaches that rely primarily on a lexicon of predefined opinion words as well as syntactic and linguistic rules.
The hybrid approach adopts a combination of the lexical approach and the machine learning approach to achieve higher accuracy [4].

3.1 Supervised approaches

As for the approach based on machine learning [5-7] proposed a novel and interesting mathematical approach to classifying message-writing feelings in Modern Standard Arabic (MSA). These functions are classified using SVM and optimized using KNN. The results of the experiments showed that such an approach is very interesting. [8] examined a corpus-based approach for AS tweets written in MSA and Egyptian dialects. They used standard n-gram functions and experimented with several classifiers (SVM and NB), via the Weka toolkit. The results obtained are very promising as a first step.
In [9-11], the authors applied their experiments on the OCA dataset of [12] using an internal dataset of 322 comments, including 164 positives, 136 negatives and 22 neutrals. [13] looked at the sentiment analysis in Arabic tweets with presence of dialectical words. In this article, SVM and NB classifiers were used for classification of feelings and the results showed the effect of varying the translation step. [14] used supervised learning to assign sentiment or polarity labels to tweets written in arabizi. The results obtained by this work reveal that the SVM accuracies are superior to the Naive Bayes accuracies. In [15], the authors worked on the AS of the Tunisian dialect. They used machine-learning techniques (SVM, NB) to determine the polarity of comments written in Tunisian dialect. [16] proposed an approach for classifying the Arabic comments of Algerian newspapers into positive / negative classes. For expe-

riments, two well-known supervised learning classifiers (SVMs) and Naïve Bayes (NB). The best results are obtained in terms of accuracy, both in SVM and NB, but the use of bi-gram increases the results in both models. The corpus SIAAC, gives more competitive results. [17] presented an approach to automatically classify the sentiments of arabizi messages as positive or negative. In the proposed approach, the Arabizi messages are first transliterated into Arabic. Then they automatically classify the transliterated corpus feeling superficial machine learning algorithms such as (SVM) and Naive Bays (NB) are used. The results of the simulations demonstrate the outperformance of the NB algorithm compared to all the others.

3.2 Unsupervised approaches

The unsupervised approach is based on a lexicon of feelings. Several studies have also been conducted using this technique. The authors in [18, 24] presented a new framework for the detection of feelings in Arabic tweets. The results reveal that lexicons are useful for sensing feelings and have been encouraging and open to future research. Approachin [19] was based on the lexicon to determine the vernacular analysis of the Arab "Algerian Arab" feeling. These authors mentioned in their article the main problems related to these characteristics and proposed an approach composed of four modules classified in: module of calculation of similarity of common sentences; pre-processing module; module stemming and detection of language; polarity calculation module. The experimental results thus obtained show that the system achieves good performance. [20] presented a tool for analyzing the sentiments of messages written in Algerian dialect. These authors evaluated this approach, using two lexicons annotated in feelings. The obtained results are encouraging and show continuous improvement after the completion of each step of their approach.

3.3 Hybrid approaches

The hybrid approach is to combine the methods used, one supervised approach and another unsupervised.
[21, 26] solved the difficulties of ensuring the quality of opinion in Arabic by proposing a hybrid method of approach and classification based on the lexicon using a classifier Naïve Bayes. The lexicon-based approach is executed by replacing certain words with their synonyms using the domain dictionary. The classification task is performed by the Naïve Bayes classifier to rank opinions based on the polarity of the positive or negative feeling. [22] described the iLab-Edinburgh Sentiment Analysis system. The system employs a hybrid approach of supervised learning and rule-based methods to predict a feeling intensity score (SI) for a given Arabic Twitter phrase. First, the supervised method uses a set of linear regression models formed to produce an initial SI score for each given instance of text. Second, the resulting SI score is adjusted using a set of rules that exploit a number of lexicons of feelings available to the public. The authors in [23] proposed a sentiment analysis approach for the Arabic language, which combines lexical and corpus-based techniques. The experimental results showed that the proposed hybrid approach outperforms that based on a corpus.
In order to synthesize all the presented works, we classify them in Table 1, concerning the levels of granularity (document, sentence and aspect),the used approach (Supervised, Unsupervised, Hybrid), the used Dataset, the used Algorithm (SVM, NB, Decision Tree, KNN, GA, etc.) and the studied language (MSA, Arabic dialect giving the type of the dialect).

Work	Level	Approach	Dataset	Algorithm	Language
[8]	Sentence	Supervised	Twitter	SVM, NB	MSA + Egyptian dialect
[4]	Document	Supervised	Twitter	Naïve Bayes, Decision Tree	Egyptian dialect
[9]	Document	Supervised	Aljaizira	SVM,NB,KNN	MSA
[5]	Document	Supervised	Hotel reviews	SVM	MSA
[13]	Document	Supervised	Twitter	SVM + NB	MSA + Arabic Jordanian
[14]	Document	Supervised	Twitter	SVM,NB	Arabizi Arab
[10]	Document	Supervised	Twitter	SVM, NB	Jordanian dialect
[25]	/	Supervised	Arabic Social networks	SVM + RF	Arabic language
[16]	/	Supervised	Newspapers	SVM+NB	Algerian Dialect
[27]	Document	Supervised	Social networks	NB + DT	English, MSA and Arabicdialect
[24]	Document	Unsupervised	Twitter	Based Lexicon	MSA + Saudi dialect
[18]	Document	Unsupervised	Twitter	Based lexicon	MSA, Arabic dialect
[19]	Sentence	Unsupervised	Social networks	Based lexicon	Algerian Dialect
[20]	/	Unsupervised	Social networks	Based lexicon	Algerian Dialect
[28]	Document	Unsupervised	Twitter	Based lexicon	Arabic language
[26]	Document	Hybrid	Twitter	Based lexicon+ SVM	MSA + Arabic Dialect
[21]	Sentence	Hybrid	Comments of Jordan hotels and resorts' residents.	Based lexicon +NB+SVM+KNN	Arabic language
[22]	/	Hybrid	Twitter	linear regression+ Based lexicon	Arabic language
[23]	/	Hybrid	/	/	Arabic language

4 Proposed approach

In this section, we describe our multilingual sentiments analysis approach. As shown in Figure 1, the proposed approach has several phases, which include data collection, cleaning and pretreatment, and finally sentiment analysis.

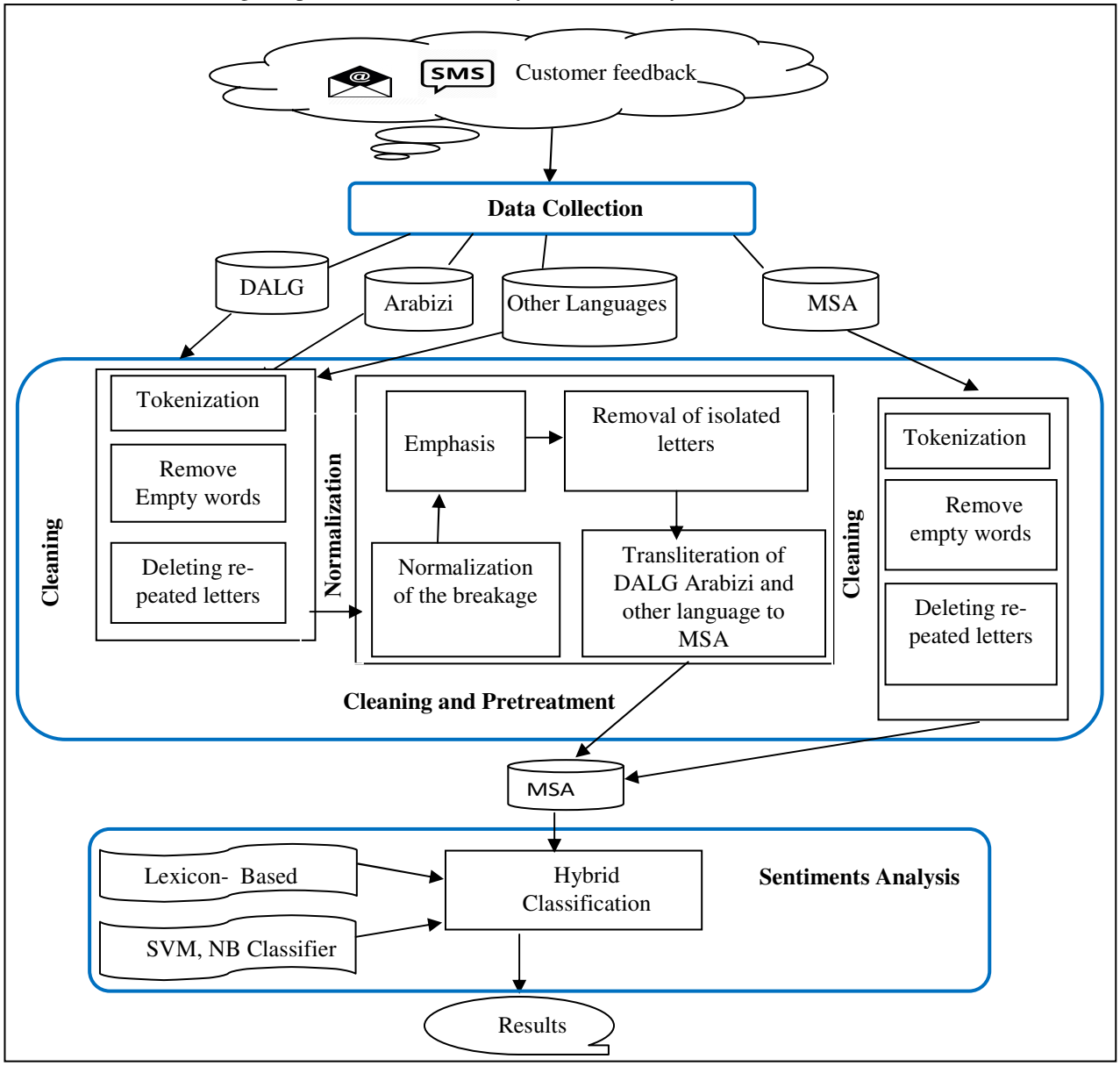

Fig. 1. General architecture of our approach

4.1 Data collection

The purpose of the collection phase is to extract data from different sources, such as files and databases. These will be stored to feed our system, aiming to target their exploitation in the analysis of feelings.
The main source of our system focus on the raw customer feedback, averaged at 20,000 comments/ month, which comes from surveys provided by the telephone company in question and is checked to determine the nature of the language used.

4.2 Cleaning and Pretreatment

Once the comments have been retrieved, they must go through the cleaning phase and only keep comments deemed useful. Indeed, the collected stream may sometimes contain duplicate messages from the same identifier.
Moreover, the system must unify the texts by the normalization of the case, the deemphasis of the words, the suppression of the single isolated letters and the maximum elimination of the noise. Indeed, we must keep only the most representative and important information. For that, it is also necessary the suppression of the empty words, as well as the suppression of the repeated letters. Then we must switch to transliteration which we translate all comments in Algerian dialect, arabizi or others in MSA
At the end of the preprocessing stage, the system must use the collected texts to perform a sentiment analysis. However, the preprocessing phase of the texts aims to eliminate a maximum of noise and to keep only the most important and representative information for our analysis. Thus, at the end of this step, we will expose the overall process to adopt and perform an effective preprocessing of the collected texts. This phase is all the more important in our work because of the context of SMS and the dialect language used. The pretreatment process involves several steps, namely:
The suppression of spaces and empty words, the normalization, the rooting, as well as the automatic translation of the Algerian dialect and Arabizi.

4.3 Sentiment analysis

To classify the customer comments of the considered operator, we will aim for the application of a hybrid approach to allow the classification of feelings by negative, positive or neutral polarity. For this, it will be considered to use the lexicon-based approach, as well as the use of supervised classification algorithms, such as Naïve Bayes (NB) and Support Vector Machines (SVM), which are selected for their efficiency and performance. Regarding the choice of the level of the analysis, we will be asked to perform the AS at the document level and this, for the following reasons, namely that comments are relatively short texts, where very often the commentary deals with only one topic and the majority of comments contain only one opinion.

5 Conclusion and perspectives

In this paper, we presented a general architecture of a sentiment analyzer for the interest of the Algerian Telephone OperatorOoredoo. The particularity of our work lies in

the proposal to build a robust system, which ensures the proper extraction and recording of data collected from the comments of the customers of the Telephone Operator. The application of sentiment analysis algorithms for several corpora, using different languages, to determine the polarity of the texts. This work is to be evaluated in an Algerian context, where the resources of Automatic Translation of Languages - TAL are poor. For future work, we will consider experimenting with the proposed solution, using both supervised and unsupervised sentiment analysis approaches.

References

1. Blog du modérateur. http ://www.blogdumoderateur.com/chifres-facebook/
2. Liu, B. : Analyse des sentiments et exploitation des opinions. In : Conférences de synthèse sur les technologies du langage humain.167pp. (2012).
3. Ahmed, F., &Nürnberger, A. : Evaluation of n-gram conflation approaches for Arabic text retrieval. Journal of the American Society for Information Science and Technology, 9(2), 1448–1465 (2009).https://doi.org/10.1002/asi.21063
4. Medhat, W., Hassan,A., Korashy,H.:Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal 5(4), 1093–1113(2014).
5. Cherif, W., Madani, A., &Kissi, M. :A new modeling approach for Arabic opinion mining recognition.InProceedings of intelligent systems and computervision (ISCV), 1–6 (2015).
6. Cherif, W., Madani, A., &Kissi, M.: A combination of low-level light stemming and support vector machines for the classification of Arabic opinions. In Proceedings of 2016 11th international conference on intelligent systems: Theories and applications (SITA)1–5(2016a).
7. Cherif, W., Madani, A., &Kissi, M.: A hybrid optimal weighting scheme and machine learning for rendering sentiments in tweets. InternationalJournal of Intelligent Engineering Informatics, 4(3–4), 322–339(2016b).
8. Shoukry,A.,Rafea,A. : Sentence-level Arabic sentiment analysis, In the International Conference on Collaboration Technologies and Systems(CTS), 2012, pp. 546–550.
9. Duwairi, R. M., &Qarqaz, I. :Arabic sentiment analysis using supervised classification. Proceedings of 2014 international conference on future internet of things andcloud (FICLOUD)579–583. IEEE,(2014).
10. Duwairi, R. M., &Qarqaz, I.: A framework for arabic sentiment analysis using supervised classification. International Journal of Data Mining, Modelling and Management, 8(4), 369–381 (2016).
11. Duwairi, R. M., Marji, R., Sha'ban, N., &Rushaidat, S. :Sentiment analysis in arabic tweets. In :Proceedings of 2014 5th international conference on information andcommunication systems (ICICS)1–6(2014).
12. Rushdi-Saleh, M., Martín-Valdivia, M. T., Ureña-López, L. A., &Perea-Ortega, J. M. :Oca: Opinion corpus forarabic. Journal of the American Society for Information Science and Technology, 62(10), 2045–2054(2011b).
13. Duwairi, R. M. :Sentiment analysis for dialectical arabic. In Proceedings of 2015 6th international conference on information and communication systems (ICICS)166–170(2015).

14.Duwairi, R.M., Alfaqeh, M.,Wardat, M., Alrabadi, A.: Sentiment analysis for arabizi text. In: 2016 7th International Conference on Information and Communication Systems (ICICS), 127–132. IEEE (2016)
15.Medhafar, S., Bougares, F., Esteve, Y., Hadrich-Belguith, L.: Sentiment analysis of tunisian dialects: Linguistic ressources and experiments. In: Proceedings of the Third Arabic Natural Language Processing Workshop. pp. 55-61 (2017).
16. Rahab, H., Zitouni, A., and Djoudi, M.: Siaac: Sentiment polarity identification on arabicalgerian newspaper comments. In Proceedings of the Computational Methods in Systems and Software, pp. 139–149. Springer(2017).
17. Guellil, I., Azouaou, F., Abbas, M., &Fatiha, S. :Arabizi transliteration of Algerian Arabic dialect into Modern Standard Arabic. Paper presented at the Social MT 2017/First workshop on Social Media and User Generated Content Machine Translation (2017a).
18.Duwairi, R., Ahmed, N.A., Al-Rifai, S.Y.: Detecting sentiment embedded in Arabicsocial media–a lexicon-based approach. J. Intell. Fuzzy Syst. 29(1):107–17 (2015).
19. Mataoui, M., Zelmati, O., Boumechache, M.: « A proposed lexicon-based sentiment analysis approach for the vernacular Algerian Arabic », Research in Computing Science, vol. 110, p. 55-70 (2016).
20. Guellil, I., Azouaou, F., Semmar, N. : Une approche fondée sur les lexiques d'analyse de sentiments du dialecte algérien : Association pour le Traitement Automatique des Langues
21. Khalifa, K., Omar, N.: « A hybrid method using lexicon-based approach and naive Bayesclassifier for Arabic opinion question answering », Journal of Computer Science, vol. 10,no 10, p. 1961(2014).
22 .Refaee, E.,Rieser,V. :iLab-Edinburgh at SemEval-2016 Task 7: Ahybridapproach for determining sentiment intensity of Arabic Twitter phrases,Proc. SemEval, pp. 474–480 (2016).
23.Biltawi, M., Al-Naymat, G., &Tedmori, S. : Arabic sentiment classification: A hybrid approach. Proceedings of 2017 international conference on New trends incomputing sciences (ICTCS)104–108(2017).
24. Albraheem,L., Al-Khalifa,H.S. : Exploring the problems of sentiment analysis in informal Arabic, In Proceedings of the 14th InternationalConference on Information Integration and Web-based Applications &Servicespp. 415–418 (2012).
25. Mustafa, M., AlSamahi, A., &Hamouda, A. :New avenues in arabic sentiment analysis. International Journal of Scientific & Engineering Research, 8(2), 907–915 (2017).
26. Aldayel, H.K., Azmi,A.M.: Arabic tweets sentiment analysis–a hybrid scheme. Journal of Information Science 42(6), (2015).
27. Elhag, M. Nordiana, A. K., Vimala, B., Ahmed,A.:Sentiment analysis algorithms: evaluation performance of the Arabic and English language. In the 2018 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE) 978-1-5386-4123-1/18/$31.00 2018 IEEE
28. Abd-Elhamid, L., Elzanfaly, D., &SharafEldin, A. :Arabic feature-based level sentiment analysis using lexicon-based approach. Journal of Fundamental and Applied Sciences, 10(4S), 143–148(2018).