ARAACOM: ARAbic Algerian Corpus for Opinion Mining*

Hichem Rahab
Laboratoire ICOSI
Université de Khenchela
Algeria
rahab.hichem@univ-khenchela.dz

Abdelhafid Zitouni
Laboratoire LIRE
Université de Constantine 2
Algeria
Abdelhafid.zitouni@univ-constantine2.dz

Mahieddine Djoudi
Laboratoire TechNE
Université de Poitiers
France
mahieddine.djoudi@univ-poitiers.fr

ABSTRACT

Nowadays, it is no more needed to do an enormous effort to distribute a lot of forms to thousands of people and collect them, then convert this from into electronic format to track people opinion about some subjects. A lot of web sites can today reach a large spectrum with less effort. The majority of web sites suggest to their visitors to leave backups about their feeling of the site or events. So, this makes for us a lot of data which need powerful mean to exploit. Opinion mining in the web becomes more and more an attracting task, due the increasing need for individuals and societies to track the mood of people against several subjects of daily life (sports, politics, television,...). A lot of works in opinion mining was developed in western languages especially English, such works in Arabic language still very scarce. In this paper, we propose our approach, for opinion mining in Arabic Algerian news paper.

CCS CONCEPTS

•Information systems~Sentiment analysis • Computing methodologies~Natural language processing

KEYWORDS

Natural Language Processing, Sentiment Analysis, Opinion Mining, newspaper, Arabic comments, machine learning.

ACM Reference format:

2017. ARAACOM: ARAbic Algerian Corpus for Opinion Mining. In *Proceedings of International Conference of Computing for Engineering and Sciences, Istanbul, Turkey, July 2017 (ICCES '17), 5 pages.*
DOI: 0.1145/3129186.3129193

ICCES '17, Istanbul, Turkey

DOI: 0.1145/3129186.3129193

1 INTRODUCTION

Arabic is the language of Holy Quran. It is also one of the major languages of the United Nations. It is the mother language of more than 330 million people in more than 22 countries[1].
We can find three types of Arabic, according to the scope of use, Classical Arabic (CA) that is conserved principally bay the holy Coran and traditional islamic literature, Modern Standar Arabic, which is the nowadays arabic used in literature, news, official correspondences in most Arabic countries, and Arabic dialects, which are spoken version of arabic whithout realy a written form[12]. In this latter form we can found two families, eastern dialects and maghrebine dialects.
The widespread of Internet and its applications, make it necessary to deal with data of several sources, and one of the most important feature of these features is the language's one.
In most applications, one needs to know what object or features of the object the opinions are on. However, the two sub-tasks of the sentence-level classification are still very important because (1) it filters out those sentences which contain no opinion, and (2) after we know what objects and features of the objects are talked about in a sentence, this step helps to determine whether the opinions on the objects and their features are positive or negative[6].
The current research is focusing on opinion target identification, especially the acquisition of vocabulary specifying a positive or negative opinion from Arabic on line press (case study the Algerian daily echorouk and elkhabar).
Unlike product evaluations, where reviews target product and its features or its parts to express positive or negative sentiment, journal reviews comments have several targets, in important part, they are not related to article topics but express someone feeling against person, events...etc, (e.g. political system, life conditions…) such comments can express positive (resp. negative) sentiment without necessary meaning what we want extract evaluation for. An important task was to recognize such off-topic comments to differentiate them from on-topic ones.
The paper is organized as fellow, in the second section related works were presented, our approach will be detailed in section three. In section four we will give our experimental evaluation nd achieved results, in section five we present our conclusion and perspectives for future works.

2 RELATED WORKS

Opinion mining domain knows a great progress in last few years. Several axes were tackled by researchers, in different languages. However, for the Arabic language there is only few achieving works.
Aila R in [11] whose the work is for Spanish, and was reproduced by Farek et Tlili [5] for Arabic, the work is based on opinion representation as 4 elements object through following conceptual model:

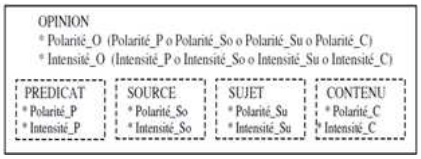

Figure 1: Conceptual Opinion Representation Model.

Where, Predicat was the principal element in which the model is based on for the opinion text segments identification. Source, or opinion holder [6], is the entity (person, …) to which opinion can be attributed. Subject, is the target of opinion, Content, is the content of opinion[11].
Then authors in [5], use a set of labels to indicate the presence of opinion on text with XML type annotation. For a given text, the processing goal was to incorporate labels for opinion and their components. The authors mark some difficulties in their work, such subject identification for opinions, which is the most important point we deal with in this work.
Based on their early work[5], In [4] the authors use domain ontology to identify comments subject, which is an object (e.g. product) or one of its features, that corresponds to an ontology O concept or a property. The system extracts Elementary Discourse Units (EDU) using delimiters such comma ",", periods, ".", "?", "!". Then in parallel, using a semantic lexicon, Elementary Opinion Units (EOU) e.g. "Excelent", "not good" are extracted, and using Domain ontology, objects features will be extracted. The result of the two parallel processes will be associated to be used in classification step.
In DOMA approach [2] authors was interested in automatic adjectives dictionary creation, which integrates domain knowledge. They used the seed list defined in [13] That contained 7 positive words P={good, nice, excellent, positive, fortunate, correct, superior}, and 7 negatives N={bad, nasty, poor, negative, unfortunate, wrong, inferior}. For each positive (resp. negative) seed word, a search engine is used with a query specifying application domain d, the seed word with the sign "+"and the negative (resp. positive) seed words with the sign "-". For each seed word a number k of documents was collected. In final, 14 corpus, 7 positive and 7 negative will be obtained to be used as training corpora. The result corpora will be used to enhance the two seed words sets. The classification step is summarized to count, for each document to classify, the number of found positive adjectives and negative ones, then if the difference is positive (resp. negative) the document is classified as positive (resp. negative), otherwise it is considered as neuter.
The authors in[8] are interested on opinion texts classification in French language, SYBILE the proposed approach is an hybrid method which combine techniques of symbolic method based on syntaxico-semantic analyzer rules and statistic method based on machine learning techniques.
The work of [3] deal with an under resources language, the Algerian Dialect, in this work the most important features of this language are introduced. And that the lack of resources for this language, an in home corpus of 6400 sentences is created by transcribing conversations of everyday life, also some TV shows and movies are transcribed. For G2P converter the authors report a correctness of 85%. And they achieve a 69% of accuracy in the evaluation with an ALG corpus.

3 OUR APPROACH

In this step we will present our proposed approach for identification of opinionated sentences in Arabic comments. **Error! Reference source not found.** below present the approach general process.

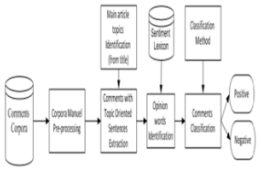

Figure 2: Our approach general process.

3.1 Corpora Manual Pre processing

Natural languages contain inherent ambiguities, and writing systems often amplify ambiguities as well as generate additional ambiguities [10].
Comments are collected from on line Arab journals, and are mostly written in Algerian Dialect (AD)" "الدارجة الجزائرية" and Modern Standard Arabic (MSA), and a little in french language. This reality make such comments processing more complicate from the reason that there are no available sentiment lexicon for this spoken language, and the most of works found in literature deal principally with Modern Standard Arabic. Also punctuation marks like comma, periods, interrogation points, exclamation points, are very less used which complicate comments sentence segmentation.
For this, a pre-processing step is strongly recommended to well dealing with such corpora.
Our preparation consists in adding punctuation marks comma, periods, interrogation points, exclamation points (e.g. in this passage "نقول للرئيس المدير العام هل هذا منطقي" an interrogation point "؟" is very useful). Correction of orthographic errors, these errors are very occurred (such us "بخابر نفطال" and the correct "مخابر نفطال"). Substitution of some spoken Algerian language words in theirs Arabic equivalent words ("يكنسيمي" substituted by "يستهلك").

3.2 Main Article Topics identification

In this step we start by extraction the main article topics, which are the comments target. This step led us, later, to identify off-topic comments.

Definition1: A *Topic T* is an entity which can be a product, person, event, organization, or object. It is associated with a set of *components* (or *parts*), *sub-components*, and so on. A Topic can be represented with a finite set of features, $F = \{f1, f2, ..., fn\}$, which includes all above parts and the object itself as a special feature[6].

Example: in the following article of an Algerian newspaper titled "نفطال", اطمئنوا.. لا غشّ في تركيبة الوقود وهذا دليل, ", which mean "*be assure … there are no frauds in carburant composition and this is Naftal's proofs* ", we can extract two main topics which are "الوقود" (carburant) and "نفطال" (Naftal, an Algerian governmental company).

We do this extraction manually, i.e. for each article we must deal with; we take the title and try to find its one or two main topics.

The importance of features of a topic reside in the fact that they can be used by reviewers to express opinions instead the topic itself, and we must recognize such substitutions.

We perform then a manual construction of features table, which will contain different expressions such synonyms, variants …etc used for a topic.

Example: for above example, we can construct a table such bellow:

Table 1: Example of Topics Features

Topics	Carburant "الوقود"	Naftal "نفطال"
Features	(essence) بنزين	(Sonatrach)سوناطراك
	(Gaz oil)مازرت	(Responsible)مسؤول
	(Composition) تركيبة	(State) الدولة
	(Concentration) تركيز	(Government) الحكومة
	(car) سيارة	(Authority) السلطة
		(Political system)النظام

3.3 Comments with Topic Oriented Sentences Extraction

Opinions are modeled as 4 elements object [5] as shown in **Error! Reference source not found.**.

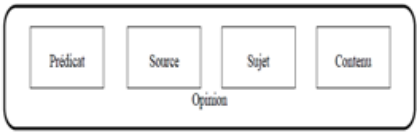

Figure 3: Opinion Representation Model.

As mentioned above, we start from article title where we extract manually the main topics of article which led us in a later step to identify opinion passages on topics or topic features, and only these passages are treated to identify subjective ones.

Definition (opinion passage on a feature): An opinion passage on a feature f of a Topic T evaluated in comment is a group of consecutive sentences in comment that expresses a positive or negative opinion on f.[6]

Example: in the passage " لقد لاحظت هذه الظاهرة المتعلقة بنفاذ البنزين بسرعة فائقة"", one of features "بنزين" (Carburant) is subject of opinion. And the opinion segmentation is shown in the table below:

Table 2: Example of opinion passage segmentation

	Arabic passage	*English mining*
	لقد لاحظت هذه الظاهرة المتعلقة بنفاذ البنزين بسرعة فائقة	I have remarked this phenomenon of essence speedily voidance
Predicate	لاحظ	have remarked
Source	ت	I, in Arabic morphology the subject can be attached to the verb
Subject	البنزين	Essence
Content	نفاذ بسرعة فائقة	speedily voidance

Then, we keep from comment only opinion passages which contain the important amount of sentiment orientation information.

3.4 Opinion words identification

A sentimental lexicon is used as input to this step, this lexicon is manual prepared from positive and negative words. Lexicon must be domain dependent, because almost approaches are confronted with the fact that the same opinion word (generally adjective) is positive for domain and negative or neuter for another domain [2].

We have achieved a simple stemming step to minimize comment vectors in classification step. Negation is used to enhance positive list.

Table 3: A sentiment Lexicon Extract

Negative words	*Positive words*
؟	جيد
!!	تطوير، تنمية
!	ليس مغشوش
تفقر	ليس مشكل
خاسرة	
خاسرة	
سريع الاستهلاك، تستهلك كثيرا، يضيع هباءا،	

سرعة نفاد، سريع
الاحتراق
خدعة، خداع، خداعهم
استخفافا، استغباء، لا
يستحون
مغشوش، يغشونا
كذبا، الكذب، بهتان
أقبح
ذنب
Yefdahkoum
فضائح
الفساد
تتعطل
ينفذ بسرعة، نفاذ بسرعة
سذج
الجحيم

4 Experimental Evaluation

The last phase of this work consists on comments classification into positive and negative classes.

We used a vector space model to represent the comments in the corpus. In the vector space model, each comment is represented as a vector in an n-dimensional space, where n is the total number of opinion words in sentiment lexicon, The result is a C*n document term matrix, where *C* is the number of comments and n is the number of words in sentiment lexicon [7].

4.1 Word vector creation

For word vector creation we use three model:

4.1.1 Term Occurrence. In this model all occurrences of a term in a document (comment) are computed and considered.

4.1.2 Term Frequency. In this model we consider the ratio of the number of occurrences of a term aver the total number of terms.

4.1.3 Binary term occurrences.

4.2 Evaluation method

For evaluation of our model we use the well known classification method Naïve Bayes, which is basic on the bayes assumption[9].

4.3 Evaluation Results

In the evaluation we use two models, Uni-gram model and Bi-gram model.

Table 4 show the results of test with Uni-gram model, it is very clear that we achieve a very high precision, come to 97.50% in the case Binary Term Occurrence model.

Table 4: Evaluation with Uni-gram using Naive Bayes

	recall	precision	**accuracy**
Term Occurrence	68,37%	**93,00%**	80,05%
Term Frequency	69,17%	**95,00%**	79,75%
Binary Term Occurrence	69,17%	**97,50%**	80,93%

In Table 5, we can see that the results are improved using the bigram model. Also the BTO (Binary Term Occurrence) model achieve the best results of 100%.

Table 5: Evaluation with Bi-gram using Naive Bayes

	recall	precision	**accuracy**
Term Occurrence	69,57%	**97,50%**	80,93%
Term Frequency	69,17%	**96,50%**	80,33%
Binary Term Occurrence	69,17%	**100%**	81,78%

For this evaluation we can conclude that the bi-gram model constitute an improvement in the work.

5 Conclusion and future works

In this work we present ARAACOM (ARAbic Algerian Corpus for Opinion Mining) which is our approach for opinion target identification, especially opinion vocabulary extraction, in the case of Arabic opinion classification on positive and negative classes.

The researches found in literature for Arabic opinion mining are rare comparing with English language for example, and within this Arabic works, we don't found ones dealing with Arabic spoken languages such as Algerian dialect.

After browsing main progress in domain researches, we present our approach general process. We prepared data as vectors, then we use Weka toolkit for data clssification. We used NaiveBayes supervised classifier to classify corpora comments into two classes.

In the evaluation we use a well known evaluation method, Naïve Bayes, which is based on the Bayes assumption. For evalution three models of vector were presented. And both, uni-gram and bi-gram were evaluation. The bi-gram model increase results with 2.5% which is an important step in this evaluation.

As perspectives of this work we would use a more general and large sentiment lexicon to deal with different subjects comments. Also punctuation is very important especially exclamation and interrogation ponits, which merit more importance in classification task. We remark also in our training corpus that 75% of comments are negative and we relate this to political nature of corpora we use, this need a more study in further works.

The use of bigram model increase considerably results, and s perspective we would in future works test the use of other model like Tri-gram model.

Manual created lists such sentimental lexicon and Topics features list can be enriched by search using similarity measures such as PMI (Pointwise Mutual Information).

REFERENCES

[1] Cherif, W, Madani, A., & Kissi, M. (2015). Towards an efficient opinion measurement in Arabic comments. Procedia Computer Science, 73, 122-129.
[2] Harb, A., Dray, G., Plantié, M., Poncelet, P., Roche, M., & Trousset, F. 2008. Détection d'Opinion : Apprenons les bons Adjectifs. Atelier FODOP, 59–66.
[3] Harrat, S., Meftouh, K., Abbas, M., Hidouci, K. W., & Smaili, K. 2016. An Algerian dialect : Study and Resources. International Journal of Advanced Computer Science and Applications, 7, March, 384–396.
[4] Lazhar, F. and Tlili-Guiassa, Y. 2012. Identification of opinions in arabic texts using ontologies, Workshop on Ubiquitous Data Mining.
[5] Lazhar, F. and Yamina, T. 2008. Identification d'opinions dans les journaux Arabes. Doctoral dissertation, Université Badji Mokhtar de Annaba.
[6] Liu, B. 2010. Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing, 2,627-666
[7] Mahyoub, F. H., Siddiqui, M. A., & Dahab, M. Y. 2014. Building an Arabic Sentiment Lexicon Using Semi-supervised Learning. Journal of King Saud University - Computer and Information Sciences. 26, 4, 417–424.
[8] Maurel, S, Curtoni, P, and Dini, L, 2008. L'analyse des sentiments dans les forums. Atelier Fouille des Données d'Opinion
[9] McCallum, A. and Nigam, K. 1998. A Comparison of Event Models for Naive Bayes Text Classification. AAAI/ICML-98 Workshop on Learning for Text Categorization, 41–48.
[10] Palmer, D. 2010. Text Preprocessing. Handbook of Natural Language Processing. 9–30.
[11] Rosá, A.2008. Identification automatique de marques d'opinion dans des textes, JEP-TALN-RECITAL '08, 9–13.
[12] Shaalan, K. 2014. A Survey of Arabic Named Entity Recognition and Classification. Computational Linguistics, 40, 2, 459-510.
[13] Turney, P.D. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. 40th annual meeting on association for computational linguistics, 417-424.