SIAAC: Sentiment Polarity Identification on Arabic Algerian Newspaper Comments

Hichem Rahab[1(✉)], Abdelhafid Zitouni[2], and Mahieddine Djoudi[3]

[1] ICOSI Labs, University of Khenchela,
BP 1252 El Houria, 40004 Khenchela, Algeria
rahab.hichem@univ-khenchela.dz
[2] Lire Labs, Abdelhamid Mehri Constantine 2 University,
Ali Mendjli, 25000 Constanine, Algeria
abdelhafid.zitouni@univ-constantine2.dz
[3] TECHNE Labs, University of Poitiers,
1 rue Raymond Cantel, 86073 Poitiers Cedex 9, France
mahieddine.djoudi@univ-poitiers.fr

Abstract. It is a challenging task to identify sentiment polarity in Arabic journals comments. Algerian daily newspapers interest more and more people in Algeria, and due to this fact they interact with it by comments they post on articles in their websites. In this paper we propose our approach to classify Arabic comments from Algerian Newspapers into positive and negative classes. Publicly-available Arabic datasets are very rare on the Web, which make it very hard to carring out studies in Arabic sentiment analysis. To reduce this gap we have created SIAAC (Sentiment polarity Identification on Arabic Algerian newspaper Comments) a corpus dedicated for this work. Comments are collected from website of well-known Algerian newspaper Echorouk. For experiments two well known supervised learning classifiers Support Vector Machines (SVM) and Naïve Bayes (NB) were used, with a set of different parameters for each one. Recall, Precision and F_measure are computed for each classifier. Best results are obtained in term of precision in both SVM and NB, also the use of bigram increase the results in the two models. Compared with OCA, a well know corpus for Arabic, SIAAC give a competitive results. Obtained results encourage us to continue with others Algerian newspaper to generalize our model.

Keywords: Opinion mining · Sentiment analysis · Arabic comments · Machine learning · Natural Language Processing · Newspaper · Support Vector Machines · Naïve Bayes

1 Introduction

The proliferation of Internet use and application in our daily life we offer a large amount of data of several forms and about all domains, this treasury need a powerful means to take benefit from. A lot of available data in the web is constituted by user generated content (UGC) like product reviews, and comments submitted by users of Web sites

R. Silhavy et al. (eds.), *Applied Computational Intelligence and Mathematical Methods*,
Advances in Intelligent Systems and Computing 662, DOI 10.1007/978-3-319-67621-0_12

such as Epinions.com and Amazon.com [1], also in websites of Algerian newspaper such Echorouk[1], elkhabar[2],… etc., or television channels like Aljazeera[3] [2].

Sentiment analysis or opinion mining [3, 4], is a new field in the cross road of data mining and Natural Language Processing/Natural Language Understanding (NLP/NLU) [5] which the purpose was to extract and analyze opinionated documents and classify it into positive and negative classes [6, 7], or in more classes such as in [8] and [9]. Unlike data mining, where the work is to track meaningful knowledge from structured data, in sentiment analysis it is subject to find structured knowledge from unstructured amounts of data [10].

A challenging task in opinion mining is comments classification to their positive and negative sentiment toward article subject. This problem will be increased in the case of Arabic language due to the morphologic complexity and the nature of comments, for instance the behavior of the reviewers could be affected by the culture in Arabic countries [7].

We have organized this paper as fellows: In the second section related works are presented, Third section of the paper deals with the approach we carried out for sentiment analysis, and the different steps we fellow are detailed. Experimental evaluation techniques and metrics are explained in the section four. The fifth is dedicated to present and discuss obtained results. We conclude the work in the sixth section with giving perspectives to future works.

2 Related Works

An important baseline to conduct studies in opinion mining is the language resources, in [7] the OCA, a publicly available corpus, is designed to implement sentiment analysis applications for Arabic language. The authors collect 500 movie reviews from different web pages and blogs in Arabic, and they take benefit from the rating system of websites to annotate them as positive or negative. For experiments two machine learning algorithms, SVM and NB were used, and their performances are compared. A 10-fold cross validation method was implemented. The best results were obtained with SVM, and the use of trigram and bigram overcome the use of unigram model.

In their next work [11] the authors of OCA [7], using a Machine Translation (MT) tool, have translated the OCA corpus into English, generating the EVOCA corpus (English Version of OCA) contains the same number of positive and negative reviews. Following the work in [7] Support Vector Machines (SVM) and Naïve Bayes (NB) were applied for classification task. Obtained results are worse than OCA (90.07% of F_measure).

The authors in [6], in the goal of improving accuracy of opinion mining in Arabic language, they investigate the available OCA corpus [7], and they use the two well know machine learning algorithms SVM and NB with different parameters of SVM.

---

[1] http://www.echoroukonline.com.

[2] http://www.elkhabar.com.

[3] http://www.aljazeera.net.

Then 10, 15, and 20 fold cross validation were used. For SVM method the highest performance was obtained with Dot, Polynomial, and ANOVA kernels. For NB, its highest accuracy was achieved with BTO (Binary Term Accuracy).

In [8] and [9] a classification in five classes with SVM method was used. In [8] Arabic reviews and comments on hotels are collected from Trip Advisor website and classified into five categories: "ممتاز" (excellent); "جيد جدا" (very good); "متوسط" (middling); "ضعيف" (weak) and"مروع" (horrible), the modeling approach combined SVM with kNN provides the best result (F_measure of 97%). The authors in [9] proposed a three steps system consists of: corpus preprocessing, features extraction, and classification. The corpus used is obtained from Algerian Arabic daily Newspapers. So they focus on the second step where 20 features were used, and a combination of many SVM was used to classify comments into five classes.

The work in [12] investigate in the sentence and document levels. For the feature selection they start by the basic known feature model, the bag-of-words (BOW), where the feature model contain only the available words as attributes. A second type of feature model is created by adding the polarity score as attribute, using SentiWordNet via a machine translation step.

The work in [13] focus on Arabic tweets to study the effect of stemming and n-gram techniques to the classification process. Also the impact of feature selection on the performance of the classifier is studied. Support Vector Machines (SVM), Naïve Bayes, (NB), and K-nearest neighbor (KNN) are the used classifiers. We remark that the authors don't study the effect of changing parameters of different classifiers. In the results the authors mention that the use of feature selection technique improves significantly the accuracy of the three classifiers, and the SVM outperforms the other classifiers.

In [2] the authors used two available Arabic Corpora OCA created in [7] and ACOM which is collected from the web site of Aljazeera channel[4]. For the classification task, Naïve Bayes, Support Vector Machines and k-Nearest Neighbor were used. Stemming is investigated and the work concludes that the use of light-stemming is better than stemming. Obtained results show that the classification performance is influenced by documents length rather than the data sets size.

It is clear from this study of related works that publicly available resources for sentiment analysis in Arabic language are seldom. And those available are generally collected from movie reviews, which limit their domain of use. This fact makes it very important, for us, to create our proper corpus to be adequate with the purpose of our study.

3 Our Approach

In this step we will present our proposed approach for sentiment analysis in Algerian Arabic Newspaper. In our work we use RapidMiner tool kit[5], which is free for educational purpose. And we collect comments from the Algerian daily echorouk[6]. Figure 1 show the general approach we adopt.

[4] www.aljazeera.net.

[5] https://rapidminer.com/.

[6] www.echoroukonline.com.

3.1 Corpus Generation

We have created our corpus for Algerian Arabic SIAAC (Sentiment polarity Identification on Arabic Algerian newspaper Comments) which mean in arabic (سياق, Context).

So we construct our corpus principally from the web site of echorouk newspaper[7]. The articles cover several topics (News, politic, sport, culture).

Compared to the important visitors of Algerian Newspaper websites, except Echorouk web site the number of comments is very low. And a lot of them are out of the main topic of the article, these comments are considered as neutral, and in the remained comments we found that negative ones largely outnumber the positives, which make for us a challenge to have equilibrium in our corpus. Table 1 present the statistics of comments collected from Echorouk web site for different categories in SIAAC before the equilibrium between different classes. It is clear that negative ones (91) largely outnumber other categories.

Table 1. Number of comments in SIAAC before equili-brium

Positive	Negative	Neutral	**Total**
32	91	24	**147**

Another difficulty is lied to the rating system, unlike review web sites such Amazon[8] for example where user can give points in scale e.g. from 0 to 5, to the article. In our case with web site Echorouk (and in the other Algerian newspaper websites) the user rather than giving points to the article, he can give one positive or negative point to other comments, this make annotation task very difficult, because having a certain number of points for a given comment have no meaning for positive or negative sentiment of the comments. So we must read carefully each comment and understand if it present a positive or negative sentiment, or even is off topic comment [14].

To generate our corpus, we take 92 comments, 60 comments from the negative category and all the 32 from the positive category (this is due to the nature of available comments -as mentioned above-) and this to have equilibrium in our corpus. The neutral comments are removed from our corpus and altered to a future work. Table 2 show the statics of our corpus.

Table 2. Number of comments from different categories in SIAAC after equilibrium

Positive	Negative	**Total**
32	60	**92**

[7] https://www.echoroukonline.com.

[8] https://www.amazon.com.

Despite the law number of comments in our generated corpus, we continue our study, due to the fact the important number of tokens per comment is 36.45, and is the documents length that influence in the classification performance more than the data set size [2].

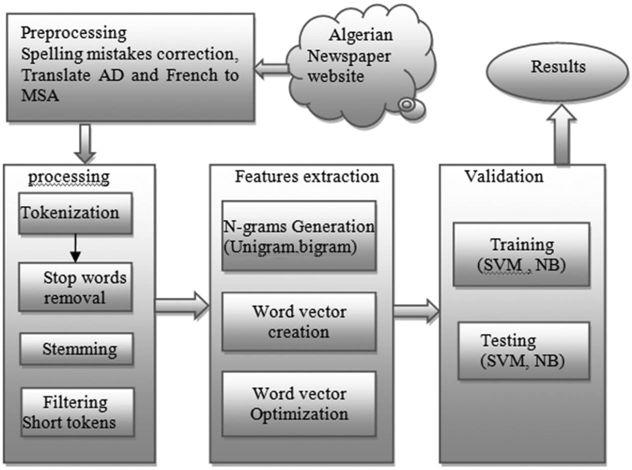

Fig. 1. Our approach general process

3.2 Pre Processing

To reduce the comment's vector size, some pre processing steps were conducted. Spelling mistakes were corrected manually, and words written in Algerian Dialect (AD) and in French are translated into Modern Standard Arabic (MSA), that stemming algorithms do not perform well with dialectical words and this dialectical words need an extended set of stopwords [15]. Characters encoding are resolved on UTF-8 (Table 3).

Table 3. Sample of comment manual correction

Extraction from Original comment	Comment after correction
عندما يقول هذا السؤول بان البنزين غير مغشوش فلماذا اصبحنا نرانه ينفذ بسرعةهذا السوال موجه اليك لانك انت بنفسك لاتراعي لهذا لانك تعبء سيارتك باطل كل شيء من عند البايلك ولاتصرف عليه مليم واحدمن جيبك اطلب من بخابر نفطال ان تفسر لنا هذا مادمت انك تقول لاوجود للغش فبماذا تفسر ذالك ياسي ايزو	عندما يقول هذا المسؤول بان البنزين غير مغشوش فلماذا اصبحنا نرانه ينفذ بسرعة هذا السؤال موجه اليك لانك انت بنفسك لاتهتم لهذا لانك تعبيء سيارتك مجانا كل شيء من الدولة ولاتصرف عليه مليم واحد من جيبك اطلب من مخابر نفطال ان تفسر لنا هذا مادمت انك تقول لاوجود للغش فبماذا تفسر ذالك ياسيد ايزو

3.3 Comments Processing

Before feature extraction, a sequence of processing steps were carried out that each comment goes through.

Tokenization
Each token representing a word, in this process we use simply spaces between words.

Stop Words removal
We note that some authors such in [2] recommend that these lists should be hand crafted as it is domain and language-specific.

Stemming
In stemming the words are reduced to their roots known as the base form or stem [16]. There are two different stemming techniques; generally stemming simply called stemming and light-stemming. For our work we use the basic Arabic stemmer.

Filtering Short tokens
Tokens with less than two letters were removed because of their low significance in opinion mining task.

3.4 Feature Selection

Feature selection is a process that selects a subset of original features to be used in the classification task [17]. The optimality of a feature subset is measured by some evaluation criterions [18]. We have used several feature selection parameters and different results are computed and compared.

N-grams Generation
In this work two n-grams were generated: Unigram and bigram.

Word vector creation
To create vector representing all comments, we use four different parameters:

1. **Term Frequency (TF):** A ratio representing the number of term occurrences over the total number of words.
2. **Term Frequency Inverse Document Frequency (TF-IDF):** Define the weight of a term in the context of a document [8].
3. **Term Occurrences (TO):** Each element represent the number the word occur in the comment (0 to n).
4. **Binary Term Occurrences (BTO):** The element takes 1 if the word appears at least once in the comment and 0 otherwise.

Word vector optimization
Very occur words and very rare ones have a low significance in opinion mining [17]. So we eliminate words appear less than 3% in corpus, and those appear more than 30%.

4 Experimental Evaluations

In this section, the proposed system is evaluated. Several experiments have been accomplished. We have used cross-validation to compare the performance of two of the most widely used learning algorithms: SVM and NB.

In our experiments, the 10-fold cross-validation has been used to evaluate the classifiers.

4.1 Classifiers

For the classification task, two well known supervised learning methods were utilized: support vector machine (SVM) and naïve Bayes (NB).

Support vector machines
The Support vector machines (SVM) are a widely used classifier in different disciplines, due to its ability to modeling diverse sources of data, their flexibility in handling data of high-dimensionality, and the high obtained accuracy.

Naïve Bayes
The Naive Bayes classifier is a well known algorithm used in text classification. In the "Naive Bayes assumption" all attributes of the examples are independent of each other given the context of the class [19]. If document belongs to different classes with different probabilities, it is classified in the class that have the highest posterior probability [13].

4.2 Performance Measures

Performance measures are defined and computed from this table; three parameters were used, precision, recall, and F_1_measure (or too simply F_mesure).

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{1}$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

Precision and recall are complementary one to the other, we combine the two using the F_1 measure called generally F_1, given as:

$$\mathrm{F_1\_measure} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{3}$$

5 Results and Discussion

5.1 Evaluation with Support Vector Machines

In the scope of our work, we have applied the SVM algorithm with two different kernel types, Anova and polynomial. And both Unigram and Bigram models were tested. For the comments vector creation four variants was implemented, Term occurrence (TO), Term Frequency (TF), Term Frequency Inverse Document Frequency (TF-IDF), and Binary Term Frequency (BTO).

Best results are obtained in precision when TF and TF-IDF are suited for vector creation and this for both SVM implemented kernels. The use of bigram increase results in most of cases (Table 4).

Table 4. Results with SVM

		Unigram			Bigram		
	Kernel	F_measure	Precision	Recall	F_measure	Precision	Recall
Term occurrence	Anova	79.64%	88.33%	72.50%	79.45%	93.33%	69.17%
	Polynomial	61.19%	47.80%	85.00%	62.93%	48.87%	88.33%
Term frequency	Anova	82.86%	96.67%	72.50%	83.50%	**100%**	71.67%
	Polynomial	67.16%	88.33%	54.17%	75.00%	**100%**	60.00%
TF-IDF	Anova	81.78%	100%	69.17%	79.39%	**100%**	65.83%
	Polynomial	70.11%	96.67%	55.00%	63.64%	**100%**	46.67%
BTO	Anova	83.00%	91.67%	75.83%	84.78%	**97.50%**	75.00%
	Polynomial	69.58%	57.40%	88.33%	68.31%	55.69%	88.33%

For comparing the results of our SIAAC corpus with OCA used in [7], we choose the term frequency vector and the anova kernel for SVM, and this because the work with OCA in [7] does not testing several SVM parameters (Table 5).

Table 5. Comparison between SIAAC And OCA in SVM classification

	N-gram model	Precision	Recall	Other metrics
SIAAC	Unigram	**96.67%**	72.50%	F_measure = 82.86%
	Bigram	**100%**	71.67%	F_measure = 83.50%
OCA in [7]	Unigram	86.99%	**95.20%**	ACC = 90.20%
	Bigram	87.38%	**95.20%**	ACC = 90.60%

The results show that in term of precision SIAAC outperform OCA, which mean that our predictive results are more important. In the other hand the recall of OCA is better than SIAAC which indicate that OCA documents are well classified than SIAAC ones.

5.2 Evaluation with Naïve Bayes

As with SVM best results are found in precision, with the different vector creation methods. And the use of bigram also increases the obtained results (Table 6).

Table 6. Results with Naïve Bayes

	Uni-gram			Bi-gram		
	F_measure	Precision	Recall	F_measure	Precision	Recall
Term occurrence	80.05(%)	95.00(%)	69.17(%)	80.93(%)	97.50(%)	69.17(%)
Term frequency	80.05(%)	95.00(%)	69.17(%)	80.93(%)	97.50(%)	69.17(%)
TF-IDF	72.15(%)	90.48(%)	60.00(%)	73.55(%)	95.00(%)	60.00(%)
BTO	80.93(%)	97.50(%)	69.17(%)	81.78(%)	100(%)	69.17(%)

As for the SVM, we will compare our the results of our corpus in classification with Naïve Bayes classifier with OCA corpus [7] (Table 7).

Table 7. Comparison between SIAAC And OCA in NB Classification

	n-gram Model	Precision	Recall	Other metrics
SIAAC	Unigram	**97.50**(%)	69.17(%)	F_measure = 80.93%
	Bigram	81.78(%)	**100**(%)	F_measure = 69.17%
OCA in [7]	Unigram	79.99(%)	**85.60**(%)	Acc = 81.80%
	Bigram	**82.75**(%)	88.80(%)	Acc = 84.60%

In this case our system SIAAC outperforms OCA in term of precision when using Uigram model, and in term of recall when using the bigram model.

6 Conclusion and Future Works

Our exploration of obvious achieved works in sentiment analysis, especially in Arabic language, show the lack in publicly available resources dedicated for carried out Arabic sentiment analysis studies. And in these rarely available ones, we found that the most are interested by movie reviews; perhaps because of the important number of web sites inciting people to review films and serials. Such available resources are not adapted to use in other domains such as newspaper comments sentiment analysis which cover several topics like politics, culture, sports, medicine, etc.

In this work we present our approach of sentiment analysis in Algerian Arabic daily newspapers. The approach starts with the corpus creation where 92 comments are collected from echorouk newspaper web site. And SIAAC (Sentiment polarity Identification on Arabic Algerian newspaper Comments) was created to be used in the scope of this study. Some processing operations were conducted.

Comments are represented in different vector models, TO, TF, BTO and TF IDF, also with unigram and bigram models. For classification, two well known methods are used, support vector machines SVM and naïve bayes NB. And different parameters for each classifier were tested. In the validation process 10-fold cross validation method was conducted to train and test the models.

Obtained results are very promising, in term of precision and recall. But in term of F_measure as a compromise between precision and recall the results remain modest which need more work to improve this rate.

Compared with the well know corpus for Arabic OCA, our approach SIAAC give a competitive results, for the SVM model SIAAC outperform OCA in classification precision. These results encourage us to continue in this issue in future works.

As perspective to this work we would to add the neutral class which allow us using all available comments (or at least an important part of them), so other methods can be implemented or a combination of existing methods can be used to resolve the multi classes problem. Also the corpus must be enriched by comments from other Algerian newspaper to generalize the model.

Another point is to dealing with comments written in Algerian dialect and French language directly without the need of the manual translation. In this point machine translation may be envisaged as an automated process.

Also it is very important to take in consideration in future works, the article topic when searching the sentiment orientation of comments which can allow us considering more comments in the classification step, that a lot of comments are removed from our corpus due to the fact that their semantic orientation is strongly related to the article topic.

References

1. Zhang, C., Zeng, D., Li, J., Wang, F.Y., Zuo, W.: Sentiment analysis of Chinese documents: from sentence to document level. J. Am. Soc. Inf. Sci. Technol. **60**(12), 2474–2487 (2009)
2. Mountassir, A., Benbrahim, H., Berraba, I.: Sentiment classification on arabic corpora. A preliminary cross-study. Doc numérique **16**(1), 73–96 (2013)
3. Liu, B.: Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol. **5**(1), 1–167 (2012)
4. Liu, B.: Sentiment analysis and subjectivity. In: Handbook of Natural Language Processing (2010)
5. Jackson, P., Moulinier, I.: Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization, vol. 5. John Benjamins Publishing Company, Amsterdam (2002)
6. Atia, S., Shaalan, K.: Increasing the accuracy of opinion mining in Arabic. In: Proceedings—1st International Conference on Arabic Computational Linguistics: Advances in Arabic Computational Linguistics ACLing 2015, pp. 106–113 (2015)
7. Rushdi-Saleh, M., Martín-Valdivia, M.T., Ureña-López, L.A., Perea-Ortega, J.M.: OCA: opinion corpus for Arabic. J. Am. Soc. Inf. Sci. Technol. **62**(10), 2045–2054 (2011)
8. Cherif, W., Madani, A., Kissi, M.: Towards an efficient opinion measurement in Arabic comments. Procedia Comput. Sci. **73**(Awict), 122–129 (2015)
9. Ziani, A., Tlili Guaissa, Y., Nabiha, A.: Détection de polarité d'opinion dans les forums en langues arabe par fusion de plusieurs SVMs. **7**, 17–21 (2013)

10. Salloum, S.A., Al-emran, M., Monem, A.A., Shaalan, K.: A survey of text mining in social media: facebook and twitter perspectives. Adv. Sci. Technol. Eng. Syst. J. **2**(1), 127–133 (2017)
11. Rushdi-Saleh, M., Martín-Valdivia, M.: Bilingual experiments with an Arabic–English corpus for opinion mining. In: Proceedings on International Conference on Recent Advances in Natural Language Processing 2011, pp. 740–745, September 2011
12. Alotaibi, S.S., Anderson, C.W.: Extending the knowledge of the Arabic sentiment classification using a foreign external lexical source. Int. J. Nat. Lang. Comput. **5**(3), 1–11 (2016)
13. Brahimi, B., Touahria, M., Tari, A.: Data and text mining techniques for classifying arabic tweet polarity. J. Digit. Inf. Manag. **14**(1), 15–25 (2016)
14. Pustejovsky, J., Stubbs, A.: Natural Language Annotation for Machine Learning: A guide to corpus-building for applications. O'Reilly Media, Inc. (2012)
15. Duwairi, R.M.: Sentiment analysis for dialectical Arabic. In: 6th International Conference on Information and Communication Systems (ICICS), 2015, pp. 166–170, February 2015
16. El-defrawy, M.: Enhancing root extractors using light stemmers. In: 29th Pacific Asia Conference on Language, Information and Computation, pp. 157–166 (2015)
17. Agarwal, B., Mittal, N.: Prominent feature extraction for sentiment analysis (2016)
18. Huan, L., Lei, Y.: Toward integrating feature selection algorithms for classification and clustering. IEEE Trans. Knowl. Data Eng. **17**(4), 491–502 (2005)
19. McCallum, A., Nigam, K.: A comparison of event models for Naive Bayes text classification. In: AAAI/ICML-98 Work Learning Text Category, pp. 41–48 (1998)