ISSN 2007-9737

Arabic Text Mining for Price Prediction of Used Cars and Equipment

Belgacem Brahimi

Mohamed Boudiaf University of M'sila,
Laboratory of Informatics and its Applications of M'sila (LIAM),
Algeria

belkacem.brahimi@univ-msila.dz, belkacem.brahimi@yahoo.fr

Abstract. Nowadays, companies and businesspersons are increasingly interested in the web for its potential and opportunities in marketing and commercial activities. Despite the importance of Internet advertising of used goods available on the web, work targeting their analysis is still limited. It is crucial for both buyers and sellers to precisely estimate the price of used products available online. Textual information that describes second hand goods is very relevant for accurate price prediction, however common solutions use only structural features for price estimation. We study in this paper the utility of using text mining techniques as well as the role of textual data integration in improving price prediction of online classifieds in Arabic. In order to evaluate the proposed methods, we collected online advertisements for two cases: used cars and lots of construction equipment. Additionally, we applied prediction algorithms to estimate the prices, namely, regression-based algorithms, K– nearest neighbor and neural network. Experimental results showed that the integration of textual features in the prediction models improves significantly price prediction compared with baseline methods that use only structured features. The results proved also that regression models are the best option for price estimation.

Keywords. Text mining, supervised machine learning, regression, used cars prices, used equipment price, price prediction.

1 Introduction

Nowadays, the Internet world is growing considerably in terms of the number of users, websites and pages. Statistics in [1] show a dramatic evolution of the Internet from 1995 until the present day (from 16 million to 5,168 million users in 2021). This increase is due mainly to the democracy of the web and the low cost of publishing and consulting web content. Consequently, the web is becoming the first solution of social, E-commerce and marketing content comprising opinions and advertisements in different domains. In last recent years, there is an increasing interest of businesspersons, companies and users in analyzing and exploiting social media content. Indeed, the capabilities of such content are various; for example, in marketing, knowing users' opinions and attitudes is very useful for both customers and companies. Another relevant case concerns the prediction of economic and commercial indicators such as incomes and prices from online textual data. Owning such knowledge could help companies to still performing and competitive in the market.

However, the rapid evolution of the huge amount of web data poses a challenge regarding its effective exploitation in an optimal time and effort. Fortunately, in text mining, automatic predictive techniques are proposed to deal with such challenging tasks as they propose solutions to analyze data and predict required outcomes by computer programs. Due to the fact that the most typical form of online information is written words, text mining has a very high commercial potential. Indeed, a study showed that 80 % of a company's data was in textual form like emails and reports [2].

Despite the utility of employing predictive text mining techniques in social media content for business purposes, there has been limited work targeting this area. The situation in Arabic is even worse because, to the best of our knowledge, no

doi: 10.13053/CyS-26-2-3185

research effort has been devoted to the application of text mining methods for the task of prediction in business and commercial domains. Actually, the state-of-the-art of Arabic text mining confirms that most research efforts have been focused mainly on thematic text categorization [3-4-5-6-7-8], sentiment analysis [9-10-11-12-13], author attribution [14-15-16-17] and mining the holy Quran [18-19-20], while other proposed papers were interested in web pages clustering and annotation [21] and information extraction [22].

Recently, and in a business perspective, many companies specialized in marketing built their websites to provide online advertisements concerning several commercial activities. The domain of used cars and equipment is among the most interesting business sectors in Algeria because of their continuous growth and expansion in the last years. In fact, the decision of the Algerian government to ban the import of new cars has revolutionized the used cars and equipment markets. For example, the famous Algerian website Oued Kniss[1], specialized in posting advertisements concerning real estates, used equipment and cars, is the first Algerian website visited in Algeria, and ranked fourth among the first international websites visited in this country in 2021 [23]. This website was worth 40 billion Algerian dinars in 2014 with more than 800,000 advertisements [24].

In contrast of new cars advertisements in which attributes are categorical representing their components and options, the description of used cars and equipment by unstructured textual data is relevant for accurate price estimation. Indeed, some textual features including car description are very valuable, and thus should be taken in consideration in price prediction. For example, the state of a given car as nearly new or good as well as its components like engine, if new, repaired or revised, affects significantly the price of used car. Table 1 and 2 illustrate two examples of car and equipment advertisements (unit U equals 10,000 Algerian dinars). In these tables, we can see some relevant textual features (in bold) that could affect the price of used cars and equipment.

The goal of this research work is to explore the capabilities of using text mining techniques in Arabic, coupled with common predictive machine learning algorithms to estimate the prices of used goods like cars and equipment. In particular, we study the influence of the text preprocessing task on price valuation. We also investigate the impact of some data mining techniques such as feature selection on prediction results. In addition, we examine and compare the performance of some predictive algorithms like K–nearest neighbor, regression-based algorithms and neural networks in price forecasting of used cars and equipment. Finally, to evaluate the contribution of using text mining and integrating textual data in the prediction model, we compare the proposed methods with the same algorithms that employ only structured variables for price prediction.

We think that proposing such solutions for predicting accurate prices will be very helpful for both buyers and sellers. Indeed, providing a precise price estimation tool allows people to make the right decision of selling or buying used properties by avoiding overestimating or underestimating the real price [25].

The rest of the paper is organized as follows. In the next section, we give an overview on some studies related to the task of prediction in the commercial context by using text mining methods. Section 3 explains the process of gathering and preprocessing textual data and describes the solution for price valuation. After this, we present the results obtained with their interpretations in section 4. Finally, section 5 provides conclusions and possible improvements of the presented work.

2 Related Work

The aim of text mining is the analysis of large amounts of textual data and the detection of linguistic usage patterns to find useful information [26]. This research area uses natural language processing and data mining techniques to extract useful knowledge from texts.

In text mining, supervised methods are among the most popular techniques for mining such valuable information. In these methods, predictive models are built and learned, and then evaluated

[1] - www.ouedkinss.com

Table 1. Sample announce of a used car with description

Car	
Model of the car	Chevrolet optra 4
Mileage in km	109,000 km
Year	2015
Description	لدي سيارة في حالة **ممتازة** فقط **خدش** صغير في الباب خلفي يمين. لا يوجد **دهن** . **محرك معاود و قوي جدا** . الباقي كل شيء **جيد**. (I have a car in **excellent condition** just a small **scratch** on the right rear door. There is no **paint**, **the engine is revised and very strong**. The rest is all **good**.)
Price of the car	140 U

Table 2. Sample announce of a lot of used equipment with description

Lot	
Lot ID	10234
Description	حصة تحتوي على شاحنة هيونداي ، حافلة ميني باص نوع كيا و سيارة بيك اب نوع تويوتا في **حالة جيدة**. (A lot contains a Hyundai truck, minibus Kia and a Toyota pick-up Toyota in **good condition**.)
Price of the lot	1410 U

Table 3. List of structured features in the car dataset

Feature name	**Type**	**Meaning**
Model	Text	Model of the car, example Chevrolet OPTRA
Mileage in km	Integer	Distance travelled by the car
Year	Integer	Year of manufacture
Price	Integer	The required price of the car

based on annotated examples with the aim of predicting the required results.

These forecasting techniques fall under two categories: classification and regression [27]. This distinction depends on the required type of prediction; in the classification task, an example is categorized in one of possible predefined classes, while regression models estimate the output of a given instance to a continuous numerical valuation [28].

In recent years, relevant research papers that investigated the task of predictive text mining models for marketing and business purposes have been proposed due to their importance for both customers and companies [25, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38]. For example, in sentiment analysis context, suggested studies attempted to analyze the impact of consumers' sentiments on company's economic outputs [29], and to estimate revenue from opinions in Social Media [30]. The effect of news headlines [31] and sentiments on stock price estimation was also investigated in the work [32]. Moreover, some scientific articles studied the relation between purchase intention and product price [33, 34].

Other interesting studies exploited text mining approaches for the task of price prediction in markets. For example, the authors [35] presented a prediction method for crude oil prices. They indicated that their approach outperforms other predictive methods. In a paper presented by [36], the researchers described a forecasting system based on text mining, called NewsCATS (News Categorization and Trading System) to predict trends in stock prices. Another interesting advertising activity in which price estimation is crucial is the real estate area. In this perspective, relevant scientific papers aimed to estimate the price of real estate classifieds [25, 37] and predict end-prices of online auctions [38] by the use of text mining methods.

Regarding Arabic text mining, research studies presented in the marketing and economic domain, that employ text mining techniques for price prediction are very rare. The present work aims to investigate this avenue of research.

3 Proposed Approach

3.1 Data Collection and Preparation

In this section, we explain the process of creating our datasets. Regarding the used cars domain, we collected Arabic advertisements from the first website of online advertising in Algeria (ouedkniss.com). The posts were gathered in the period between 10th August 2018 and 17th October 2018. We selected only texts in which web users employed standard or comprehensible Arabic in their online advertisements. Duplicate documents including identical descriptions of the second-hand cars were removed. Additionally, advertisements using Romanization of Arabic or different languages such as French were discarded from the corpus. Unrealistic announces requiring exaggerated prices were also rejected as they could affect negatively prediction model performance.

The obtained dataset contains 400 documents about the 20 most used models in Algeria. Each document in the data collection comprises two parts: the first part includes structured variables describing the used car like year of manufacture and mileage in km, while the second part of the document comprises textual data that describes the car state. The price in this dataset varies from 800,000 AD to 4.000,000 AD. AD means Algerian dinar. Table 3 shows a list of the structured features in the used car dataset.

To compile the second dataset related to used equipment of construction, we gathered examples from the same website (Ouedkniss). In addition, the same methodology of selecting documents in the first dataset of used cars is applied to create the second collection related to used equipment. In the second dataset, each document describes heavy equipment lots such as trucks and buses.

The obtained dataset comprises 482 texts, and the interval of price in this collection is between 1,000,000 AD and 28,000,000 AD.

Concerning data preparation, we performed usual text preprocessing techniques including tokenization, removing non Arabic letters and normalization of Alif – Taa and Yaa (ا- ـة- ي). In addition, stop words such as (في- على in English on-in), and words having length less than 2 letters were removed from the corpus.

In the preprocessing method, stemming is an optional task that reduces word forms to a unique representation (stem, base or root). In Arabic, widely known stemming methods are root stemming [39] and light-stemming [40]. In our work, we applied light stemming.

Regarding feature types, we used n-grams of words [41]. We recall that an n-gram of word is a contiguous sequence of n words from a given sample of text. An n-grams of size 1 is called unigrams; two successive words are called bigrams (or digrams) and size 3 is named trigrams.

The next step in the preprocessing task consists on assigning for each feature (word) a weight representing its relevance in the text. There are several weighting schemes such as Boolean weighting, Term Frequency TF and Term Frequency Inverse Document Frequency (TF.IDF) which is a combination of TF and Inverse Document Frequency (IDF). TF is the number of times a word occurs in a text, while IDF is the number of total documents over the number of documents containing word i. TF.IDF is a popular weighting scheme used in text mining applications such as information retrieval and text classification. This scheme reflects the relevance of a feature (word) in a given corpus.

Finally, we applied feature selection on the textual features to optimize the number of words in the dataset. We retained the most relevant features based on correlation between description

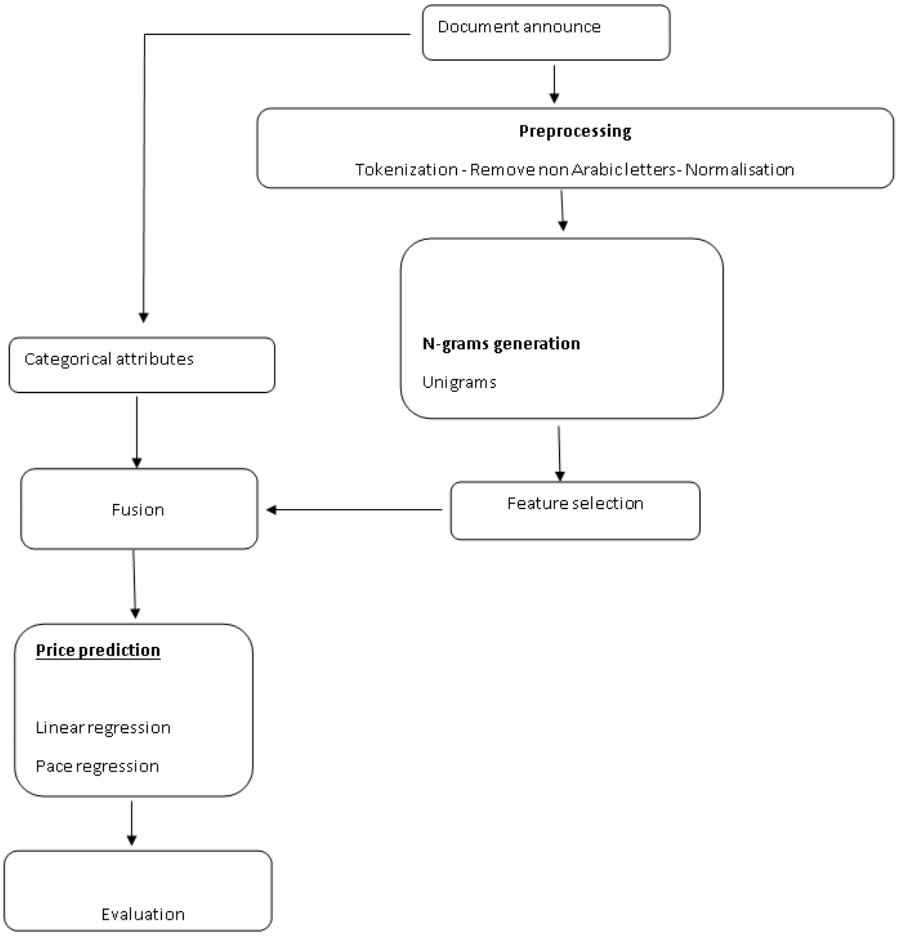

Fig. 1. Steps followed in our proposed approach

words and the output, i.e., the price. The optimal number of the most important features is determined empirically by experiments.

3.2 Used Algorithms

Regarding the price prediction model, and for comparison purpose, we applied four forecasting algorithms namely, linear regression (LinReg), pace regression (PaceReg), K-nearest neighbors (KNN) and finally Neural networks (NeuNet) for estimating the price of used cars and equipment. To determine the importance of textual information integration that includes the description of used cars in price valuation, we performed two prediction models. The first base model includes only categorical and structured data such as year of manufacture and mileage, while the second prediction model combines the first model (categorical features) with textual data that contains the description of the used car state. Figure 1 shows the different steps followed in the solution for integrating textual features to accurately predict the price of used cars and equipment.

3.3 Evaluation

To assess the performance of our predictive techniques, we calculated Root Mean Squared Error (RMSE), which is a well-known measure for evaluating prediction models. In addition, 10 fold-

Table 4. RMSE in the car dataset without textual data integration

Algorithms	RMSE
LineReg	32.791 ± 11.870
KNN (N=1)	73.631 ± 12.666
KNN (N=5)	59.165±12.021
KNN (N=10)	56.012±10.760
PaceReg	31.937±11.091
NeuNet	**30.539±10.770**

Table 5. RMSE in the car dataset with textual data integration

Algorithm	Feature size	RMSE	
		Unigrams	Bigrams
LineReg	10	27.839 ± 9.457	**27.585 ± 8.874**
	50	30.770 ± 8.798	31.414 ± 12.502
	100	30.476 ± 9.782	32.062 ± 13.055
KNN (N= 10)	10	56.106 ± 10.664	55.897 ± 10.887
	50	55.911 ± 10.881	**55.786 ± 10.728**
	100	55.907 ± 10.877	55.831 ± 10.791
PaceReg	10	**28.565 ± 8.567**	28.731 ± 8.370
	50	29.587 ± 8.996	31.228 ± 11.970
	100	32.230 ± 10.123	31.680 ± 9.079
NeuNet	10	29.136 ± 9.358	**28.526 ± 9.516**
	50	31.477 ± 9.223	29.081 ± 8.949
	100	32.397 ± 6.622	31.000 ± 8.243

cross validation was employed to average the performance results in the experiments.

Cross-validation is a popular technique used to evaluate and compare machine learning algorithms. For each iteration, we split data into two parts: one used to learn and build the prediction model, while the other part is utilized for the test step. The performance results of the 10 folds are then combined and averaged [42].

4 Experimental Results

The role of the experimental study is to evaluate and compare the performance of the proposed models for price forecasting. To perform experiments, we used the Rapid Miner software[2], which includes different tasks required to perform text preprocessing, prediction and evaluation.

Regarding the first text collection (car), we considered two scenarios. The first one is without considering textual data and using only structured features, while in the second setting, we integrated unstructured textual data representing the used car description in the prediction process.

Table 4 illustrates performance prediction results measured by (RMSE) for the four prediction algorithms without including textual information, that is the description of the car in the model. Through this table, the best result is obtained when the algorithm neural networks (NeuNet) is applied (30.539±10.770).

[2] -http://rapid-i.com

Table 6. Example of some relevant words impacting the price of the car

Word	Weight	Word	Weight
جميلة (nice)	+22.551	قوة (power)	+13.684
جديدة (new)	+55.126	قوي (powerful)	+9.627
خدوش (scratches)	-4.900	نظيفة (clean)	+6.866
ضربة (choc)	-2.898	نظافة (cleanliness)	+11.451
حادث (accident)	-15.582	نقية (clean)	+20.740
مصبوغة (painted)	-27.879	صبغة (paint)	-15.582

Table 7. RMSE in the lots of equipment dataset

Algorithm	Feature size	RMSE	
		Unigrams	Bigrams
LineReg	10	457.488 ± 238.081	518.347 ± 231.582
	50	355.422 ± 262.844	428.134 ± 260.230
	100	**318.524 ± 274.326**	345.747 ± 278.843
KNN (N= 10)	10	466.195 ± 241.201	489.468 ± 243.697
	50	398.128 ± 263.475	452.539 ± 256.848
	100	415.493 ± 263.154	**395.590 ± 269.428**
PaceReg	10	459.948 ± 244.777	511.549 ± 233.568
	50	362.932 ± 263.316	429.844 ± 255.480
	100	**340.889 ± 268.593**	366.443 ± 271.005
NeuNet	10	524.814 ± 231.400	543.280 ± 235.249
	50	**368.699 ± 257.011**	430.134 ± 251.625
	100	443.707 ± 377.139	438.771 ± 321.049

In addition, KNN is the worst prediction algorithm, and it provides the best results when the number of neighbors equals 10.

Therefore, we continue to use this value for KNN in the next experiments.

In the second step, we integrated unstructured features including the description of the car in the prediction model. For this, we tested different text representation schemes in the experiments: binary, TF and TF*IDF. We found that TF*IDF is the best for all the prediction algorithms. Hence, we provide results concerning only this weighting scheme TF*DF for unigrams and bigrams of words. In addition, we performed pruning, which eliminates words that are correlated between them in order to optimise the list of relevant features in the final list of attributes. Table 5 depicts performance results of the four algorithms when employing unigrams and bigrams. The best results for each algorithm are highlighted in bold.

From the results provided in Table 5, we see that integrating textual information that contains the description of the used cars ameliorates price prediction for all the tested algorithms compared with the results of Table 4. We can also observe that the best algorithm for predicting car price is linear regression (LineReg).

Moreover, using bigrams enhances slightly the performance in three algorithms, and the optimal number of features equals 10 yielded the lowest values of RMSE. The first algorithm gained from the integration of unstructured textual data is LineReg as RMSE was reduced from 32.791 ± 11.870 to 27.585±8.874. This result agrees with the findings of the paper [25], which confirmed that linear regression outperformed neural networks.

In order to go further in the analysis regarding the contribution of textual data integration for price estimation, we show in Table 6 a list of some relevant words that affect positively or negatively the prices of used cars.

We observe from Table 6 that some opinion words such as جديد- جميل (new and nice, in English) have a positive impact (weight) for increasing the price of the car, while some words related to the car description such as حادث – خدوش (accident and scratches in English), impact negatively its price. We see also that some words such as نقية – نظيفة (in English clean) share the same meaning. Therefore, we think that integrating a semantic approach that maps synonyms to their common concept could improve price prediction of used cars.

We continue our experiments with the second text collection of used equipment. The same preprocessing tasks were performed as in the first data collection of used cars. This data collection comprises only textual information (no categorical features).

The experimental outcomes are presented in Table 7 where the best results for each algorithm are highlighted in bold. The first remark is that the obtained results are modest when compared with the first dataset of used cars. These results go along with the conclusions of the study in [38] as the authors showed that their regression models did not provide good results for the prediction of end-prices of auction items. Our dataset is similar to the auctions dataset in [38] as it contains several heterogeneous items for sale.

According to Table 7, we see that, again, linear regression (LineReg) proves its superiority on the other algorithms in terms of RMSE, while KNN is the worst one. In addition, the optimal number of features is 100 in three algorithms. Regarding feature types, applying bigrams of words does not improve price prediction.

Finally, after carrying out our experiments on the two data collections, we have drawn some conclusions. Firstly, regarding the employed algorithms, the best ones using text mining and textual data for predicting the price are regression based models: linear and Pace regression as they provide the least RMSE values, while KNN is the worst model for price forecasting. In addition, the algorithm neural networks is applied with its defaults parameters; we think that optimizing these parameters could improve prediction results for this algorithm. Secondly, the integration of textual data that comprises the description of the car state ameliorates significantly price prediction results in the car dataset. Moreover, employing bigrams of words enhances price estimation in this corpus.

Concerning the datasets used in the experiments, the performance results of the second dataset that contains used equipment are lower than the car dataset. This is due to three factors. Firstly, each document in the equipment data collection comprises a description of many different equipment lots. In this case, the content of the lot is heterogeneous. Secondly, the description of the list of used items is not sufficient to describe them, in contrast of the car dataset, in which the description of the used car is provided in detail. Thirdly, the range of the price in the equipment dataset is larger than in the car corpus, and this obviously increases the price prediction error.

5. Conclusion and Future Work

In this paper, we studied the contribution of using text mining in Arabic for enhancing price prediction results for used cars and equipment. The idea behind this work is that textual information that describes a used good state is very relevant to precisely estimate its price. In particular, we used both types of features related to second-hand cars and equipment, namely, structured variables and textual data to makeprediction.

For this perspective, we compiled two datasets related to used cars and equipment. In addition, we employed and tested different predictive models, namely, K–nearest neighbors, regression based algorithms and neural networks to compare their performance results.

Experimental results proved that using text mining and considering textual data fusion in the prediction process improves price estimation. The results showed also that linear regression is the most suitable model for the price prediction task. As future work, we intend to apply the proposed solution on other domains. We also think that deep learning techniques, information extraction and semantic approaches would be a solution for improving prediction results.

References

1. **Internetworldstats (2021).** https://www.inter net worldstats.com/stats.htm.
2. **Tan, A.H., Ridge, K. (1999).** Text mining: The state of the art and the challenges. Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases, Vol. 8, pp. 65–70.
3. **Khorsheed, M.S., Al-Thubaity, A.O. (2013).** Comparative evaluation of text classification techniques using a large diverse Arabic Dataset. Language Resources and Evaluation, Vol. 47, No. 2, pp. 513–538. DOI: 10.1007/s10579-013-9221-8.
4. **Al-Anzi, F.S., Abu-Zeina, D. (2017).** Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing. Journal of King Saud University-Computer and Information Sciences, Vol. 29, No. 2, pp. 189–195. DOI: 10.1016/j.jksuci. 2016.04.001.
5. **Eldos, T.M. (2003).** Arabic text data mining: A root-based hierarchical indexing model. International Journal of Modelling and Simulation, Vol. 23, No. 3, pp. 158–166. DOI: 10.1080/02286203.2003.11442267.
6. **Nehar, A., Benmessaoud, A., Cherroun, H., Ziadi, D. (2014).** Subsequence kernels-based Arabic text classification. IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA), pp. 206–213. DOI: 10.1109/AICCSA.2014.707 3200.
7. **Atlam, E.S., Morita, K., Fuketa, M., Aoe, J. I. (2011).** A new approach for Arabic text classification using Arabic field- association terms. Journal of the American Society for Information Science and Technology, Vol. 62, No. 11, pp. 2266–2276. DOI: 10.1002/ asi.21604.
8. **Zubi, Z.S. (2009).** Using some web content mining techniques for Arabic text classification. Proceedings of the 8th WSEAS International Conference on Data Networks, Communications, Computers, Stevens Point, Wisconsin, USA. World Scientific and Engineering Academy and Society, pp. 73–84. DOI: 10.5555/1670344.1670357.
9. **Abdellaoui, H., Zrigui, M. (2018).** Using tweets and emojis to build TEAD: an Arabic dataset for sentiment analysis. Computación y Sistemas, Vol. 22, No. 3. DOI: 10.13053 /CyS-22-3-3031
10. **Ahmad, K., Cheng, D., Almas, Y. (2007).** Multi-lingual sentiment analysis of financial news streams. International Workshop on Grid Technology for Financial Modeling and Simulation, SISSA Medialab, Vol. 26.
11. **Mulki, H., Haddad, H., Bechikh-Ali, C., Babaoğlu, I. (2018).** Tunisian dialect sentiment analysis: A Natural Language Processing-Based Approach. Computación y Sistemas, Vol. 22, No. 4. DOI: 10.13053/cys-22-4-3009.
12. **Aldayel, H.K., Azmi, A.M. (2016).** Arabic tweets sentiment analysis – A hybrid scheme. Journal of Information Science, Vol. 42, No. 6, pp. 782–797.
13. **Al-Sallab, A., Baly, R., Hajj, H., Shaban, K. B., El-Hajj, W., Badaro, G. (2017).** Aroma: A recursive deep learning model for opinion mining in Arabic as a low resource language. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), Vol. 16, No. 4, pp. 25.
14. **Alsmearat, K., Al-Ayyoub, M., Al-Shalabi, R., Kanaan, G. (2017).** Author gender identification from Arabic text. Journal of Information Security and Applications, Vol. 35, pp. 85–95.
15. **Abbasi, A., Chen, H. (2005).** Applying authorship analysis to Arabic web content. International Conference on Intelligence and Security Informatics, pp. 183–197, Springer, Berlin.
16. **Altheneyan, A.S., Menai, M.E.B. (2014).** Naïve Bayes classifiers for authorship attribution of Arabic texts. Journal of King

Saud University-Computer and Information Sciences, Vol. 26, No. 4, pp. 473–484.

17. **Alsmearat, K., Shehab, M., Al-Ayyoub, M., Al-Shalabi, R., Kanaan, G. (2015).** Emotion analysis of arabic articles and its impact on identifying the author's gender. IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA), pp. 1–6.
18. **Sharaf, A. M. (2009).** The Qur'an annotation for text mining. Transfer report school of Computing, Leeds University.
19. **Muhammad, A.B. (2012).** Annotation of conceptual co-reference and text mining the Qur'an. University of Leeds.
20. **Alhawarat, M., Hegazi, M., Hilal, A. (2015).** Processing the text of the Holy Quran: A text mining study. International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 6, No. 2, pp. 262–267.
21. **Alghamdi, H.M., Selamat, A., Karim, N.S.A. (2014).** Arabic web pages clustering and annotation using semantic class features. Journal of King Saud University-Computer and Information Sciences, Vol. 26, No. 4, pp. 388–397.
22. **Harrag, F. (2014).** Text mining approach for knowledge extraction in Sahîh Al-Bukhari. Computers in Human Behavior, Vol. 30, pp. 558–566.
23. **Alexa (2021).** https://www.alexa.com/topsites/countries/DZ.
24. **Wikipedia (2018).** https://fr.wikipedia.org/wiki/Oued_Kniss.
25. **Abdallah, S., Khashan, D.A. (2016).** Using text mining to analyze real estate classifieds. International Conference on Advanced Intelligent Systems and Informatics, pp. 193–202, Springer, Cham.
26. **Sebastiani, F. (2002).** Machine learning in automated text categorization. ACM Computing Surveys (CSUR), Vol. 34, No. 1, pp. 1–47.
27. **Ye, N. (2013).** Data mining: Theories, algorithms, and examples. CRC press.
28. **Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. (1996).** Knowledge Discovery and Data Mining: Towards a Unifying Framework. In KDD, Vol. 96, pp. 82–88.
29. **Ghose, A., Ipeirotis, P.G. (2011).** Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. IEEE Transactions on Knowledge and Data Engineering, Vol. 23, No. 10, pp.1498–1512.
30. **Asur, S., Huberman, B.A. (2010).** Predicting the future with social media. Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01, pp. 492–499.
31. **Kirange, D.K., Deshmukh, R.R. (2016).** Sentiment Analysis of News Headlines for Stock Price Prediction. Composoft, an International Journal of Advanced Computer Technology, Vol. 5, No. 3, pp. 2080–2084.
32. **Oh, C., Sheng, O. (2011).** Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement. In Icis, pp. 1–19.
33. **Grewal, D., Krishnan, R., Baker, J., Borin, N. (1998).** The effect of store name, brand name and price discounts on consumers' evaluations and purchase intentions. Journal of Retailing, Vol. 74, No. 3, pp. 331–352.
34. **Kwun, J. W., Oh, H. (2004).** Effects of brand, price, and risk on customers' value perceptions and behavioral intentions in the restaurant industry. Journal of Hospitality & Leisure Marketing, Vol. 11, No. 1, pp. 31–49.
35. **Yu, L., Wang, S., Lai, K.K. (2005).** A rough-set-refined text mining approach for crude oil market tendency forecasting. International Journal of Knowledge and Systems Sciences, Vol. 2, No. 1, pp. 33–46.
36. **Mittermayer, M.A. (2004).** Forecasting intraday stock price trends with text mining techniques. 37th Annual Hawaii International Conference on System Sciences, pp. 10.
37. **Stevens, D. (2014).** Predicting real estate price using text mining. Department of Communication and Information Sciences. Tilburg University.
38. **Ghani, R., Simmons, H. (2004).** Predicting the end-price of online auctions.

In International workshop on data mining and adaptive modelling methods for economics and management.

39. **Khoja, S., Garside, R. (1999).** Stemming Arabic text. Computing Department, Lancaster University, UK.

40. **Larkey, L.S., Ballesteros, L., Connell, M.** E. **(2007).** Light stemming for Arabic information retrieval. Arabic Computational Morphology, pp. 221–243.

41. **Shannon, C.E. (1948).** A mathematical theory of communication. Bell System Technical Journal, Vol. 27, No. 3, pp. 379–423.

42. **Manning, C.D., Schütze, H. (1999).** Foundations of Statistical Natural Language Processing. MIT press.

Article received on 05/09/2021; accepted on 05/12/2021.
Corresponding author is Belgacem Brahimi.