ISSN 2278-3091
Volume 10, No.2, March - April 2021
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse251022021.pdf
https://doi.org/10.30534/ijatcse/2021/251022021

Feature Selection Optimization for Highlighting Opinions Using Supervised and Unsupervised Learning on Arabic Language

Khaled M. Alalayah[1,2*],Ibrahim M. G. Alwayle [1.4], Fahd A. Alqasemi[3], Nashwan A. Al-Majmar[2]
[1]Department of Computer Science, College of Science and Arts, Sharurah,Najran University, Saudi Arabia
[2] Department of Computer Science, Faculty of Science.IBB University, Yemen.
[3]Department of Information Technology, University of Science and Technology, Yemen.
[4]Faculty of Science and Arts,Amrran University, Yemen.
kmalalayah@nu.edu.sa
kh101ed2005@yahoo.com

ABSTRACT
Text mining utilizes machine learning (ML) and natural language processing (NLP) for text implicit knowledge recognition, such knowledge serves many domains as translation, media searching, and business decision making. Opinion mining (OM) is one of the promised text mining fields, which are used for polarity discovering via text and has terminus benefits for business. ML techniques are divided into two approaches: supervised and unsupervised learning, since we herein testified an OM feature selection(FS)using four ML techniques. In this paper, we had implemented number of experiments via four machine learning techniques on the same three Arabic language corpora. This paper aims at increasing the accuracy of opinion highlighting on Arabic language, by using enhanced feature selection approaches. FS proposed model is adopted for enhancing opinion highlighting purpose. The experimental results show the outperformance of the proposed approaches in variant levels of supervisory,i.e. different techniques via distinct data domains. Multiple levels of comparison are carried out and discussed for further understanding of the impact of proposed model on several ML techniques.

Key words: Natural Language Processing, Machine Learning, Opinion Mining, Sentiment Analysis, Feature selection ,NBC, KNN, K-means, Fuzzy C-means, Arabic language

.

1. INTRODUCTION

Nowadays, there are incremental interests in text mining (TM)and NLP, because TM and NLP focus on understanding textual data. Business world had a massive growth of computerized text, which appears in several contents applications, this growth is the most important reason that leads to improve NLP and TM. Available massive text needs to be utilized, and converted from stored text statistics to useful knowledge, which exhibit many benefits for business and governments [1].Opinion mining (OM), or sentiment analysis (SA)had added a lot of services to data science and textual data mining disciplines [2]. By OM, a kind of knowledge, is extracted divided into two end-decisions: Positive polarity and negative polarity[1].

Text Positive polarity is the case of writer agreement sentiment, whereas negative polarity is the writer disagreement on a text, some literature had added third polarity, which is the neutral case[2]. This paper has focused on two polarities option, as many literatures did[3][4]and [5].A text neutral polarity made it an objective text, however, in SA, they supposed to discover the polarity in a subjective text, further, in the best SA cases, neutral could be considered as weak positive or weak negative polarity. Besides, some researches worked in order to distinguish between subjective text and objective text, which is a different sentiment analysis challenging, wherein the concern is about the existence of sentiment polarity, regardless the polarity itself[1].

In this paper, we have proposed an SA feature selection approach, for enhancing opinion recognition on Arabic language text; we have tested proposed approach compared to traditional method ,that is used to be extracted in text mining experiment. Proposed main contributions are explained on the following general points:

i. Considering sentiment expressions nature and Arabic morphological specialty.
ii. Utilizing a deep scope of lexicon-based feature selection for SA.
iii. Comparing SA predictive and descriptive tasks, by the use of ML two different approaches; supervised and unsupervised learning.
iv. Comparing two Arabic text domains; an important and rarely tested Arabic SA challenge.

In SA pre-processing step, most researchers want to reduce and clean textual data records, so they begin with eliminating unnecessary terms and aggregating similar terms into less number of stems. Stemming step is done in various ways, sometimes they aggregate terms to lemma, instead of stem. Both stemming and lemmatizing processes have reduced the number of terms successfully. Although it is a reasonable process, in Arabic language it is has few feasibility with SA. In SA generally the morphological form of the word is important, for e.g. the adjective term gives more sentiment expression than other terms forms like nouns or verb, this sentiment expression

Arabic due to its rich morphological and diversity of its lexemes and lemmas. The word morphology in Arabic language is more complicated, as an example it had infix letters among the word rather than suffix and prefixes, also these affixes are not just letters changing word morphology, they are in some cases other words, so you can see a sentence in Arabic that had been written as one word, this happens because of the Clitics, which are affixes letters presented in words of other languages like English. So in pre-processing step we had focused only on the eliminating of stop-words, considering the nature of SA and the speciality of Arabic language.

Proposed feature selection mechanism is based on sentiment lexicon, however, we have proposed a different, deeply FS method, since in some researches like [3] and [4], they had exploited the sentiment lexicon to gain additional feature on traditional FS, which depended on counting the number of terms in a text record, some literature named it document. As explained in section 3, our FS proposed methodology outcome is new values related to each unique term and text record, the value is calculated using a formula giving a weight for each term, those weights are computed latter to give the values the identified relation between terms and text records.

In textual data mining, they consider supervised learning as predictive task of data, because they depend on previously known data records. To predict new ones, however, based on unsupervised learning, where there is no previously known information, it is considered as descriptive task of data[1]. We had tested the difference between two tasks, by using four techniques, two belong to predictive task, and two belong to descriptive task, the discussion in section 5 illustrated this by using graphs Figures.

Movie reviews [6]and book reviews [7]Arabic data sets are used, and compared, in our results. Also we had got the chance of testing two kinds of book review data sets, balanced and unbalanced. By balancing we mean the distribution of positive text records over negative ones, although, we had got only one balanced data set of movie reviews. movie reviews was good with long text in each record, but the number of records was less than the possibility of diving into balanced and unbalanced data sets.

The remaining sections begin with background and other related works are presented in sections 2, then we present section 3 with a proposed approach explanation, section 4 includes experiments, section 5 presents the results and discussions, finally conclusions and future works section are showed in section 6.

2. BACKGROUND & RELATED WORK:

A background overview is initiated in this section, it presents a brief introduction about techniques that are used in this study, and about core subject of the study; i.e. sentiment analysis. Some related works and a short explanation of traditional feature selection methods are presented.

2.1 Machine Learning Techniques: The four ML techniques are from both two ML approaches, i.e. supervised and non-supervised learning. Each approach has its methodology and implementation techniques.

2.1.1 Supervised Learning is the approach that trains the data on a class labelled data set, then it tests a test set and predicts later depending on the model that built during training phase, k-Nearest Neighbour and Naïve Bayes Classifier are two popular supervised learning techniques which we utilize in the experiments.

Naive Bayes Classifier (NBC): Because the learning of NBC is considered as supervised learning, we need in training phase to know the actual classes, in order to predict the classes, later in testing phase. Text mining usage of NBC is applied by the following two equations, since NBC need to find an estimation of class probability*(c)*, and word k probability is given class *c*:*P*(w_k |*c)*, this is done by:

$$P(c) = \frac{N_c}{N} \text{ and } P(w_k|c) = \frac{N_{w_k}}{\sum_{w_i \in W} N_{w_i}} \quad (1)$$

In equation (1), *N* is the total number of corpus documents, *Nc* is the number of documents that belong to class *c*, w_k one of *W* unique corpus words list, N_{wk},N_{wi} are the frequency of words w_k, w_i, respectively in class *c*. With these two estimations, the calculation of equation (2) is easy, wherein we predict the most likely class of document *d* as the following:

$$c^* = argmax_{c \in C} \left\{ \log P(c) + \sum_{k==1}^{nd} [t_k \log P(w_k|c)] \right\} \quad (2)$$

Where *c** is the predicted class for document *d* from the given classes *Cand* is the number of unique words in document *d* and t_k is the frequency of word w_k.[1]

K-Nearest Neighbour (KNN): KNN is a supervised learning classification algorithm that finds a place for each test data elements, by KNN algorithm, which starts with *k* suitable value, i.e. the number need to find nearest training data elements, in order to estimate test data probability later. Since, for *x* variables we look for those *k* nearest training set instances, usually using a distance measure then compute the function *V(x)* from following Eq. (3), then*(x)* is exploited in *p(x)*and gained by Eq. (4), the latter approximated result is helped to find the intended class of*x*[8].

$$V(x) = \pi\rho^2 \quad (3)$$

Where ρ is the distance values between *x* variables and training set furthest *k*Neighbour to those variables, *V* function outputs are a parameter in the following equation:

$$p(x) \approx \frac{k}{N \times V(x)} \quad (4)$$

Where *N* is the number of training set elements and *p* is the predicted probability of *x* which is supposed to determines class.

2.1.2 Non-supervised learning is the approach that describes data, it detects data implied aspects and gives an *N* division of the data, without previously known human labelled classes. There are many techniques implemented this approach in different algorithms, such as K-means and fuzzy c-means, we have exploited both techniques for two reasons: finding the accuracy comparison supervised technique and evaluating our feature selection approach comparison traditional feature selection methods.

K-means: K-means is a popular algorithm that is used for cluster analysis in data mining. It is used here for

discriminating features extracted comparison other feature selection methods.

K-means clustering aims to partition the *n* observations into k sets *($k \leq n$) , $S = \{S_1, S_2, \ldots, S_k\}$* so as to minimize the within-cluster sum of squares. The default and efficient distance measure in k-means is Euclidean distance, the formula used in this algorithms is [9].

$$arg_s \min \sum_{i=1}^{k} \sum_{xj \in S_i} \|x_j - \mu_i\|^2 \qquad (5)$$

Where k is the number of clusters, *x* is the j^{th} point in S dataset.

Fuzzy C-means Technique: Fuzzy c-means is a data clustering, where data elements are grouped into *c* clusters, with every data item in the data elements belonging to any cluster to certain level. A closer data item to the cluster centre has a high membership to the cluster than far one in the same cluster. It is an algorithm which finds out optimal c clusters, gains them by minimizing an objective function, and illustrated in the following:

$$J_{fcm}(U) = \sum_{i}^{c} \sum_{k}^{n} (x - v_{ik})^m . |x_k - x_i|^2 \qquad (6)$$

Where the results are centres of each cluster, C = {c_1, …, c_c} , X = {x_1, x_2, …, x_n} is finite elements set, and membership results are U= u_{ij}, which belongs to [0, 1][10].

2.2.2 Sentiment analysis and opinion mining approaches:

The work [2] was the first research which coined the terms Opinion Mining (OM) and sentiment analysis (SA). Authors presented how to extract sentiments from documents automatically. In this study, they indicate that it is essential to identify how textual sentiments are expressed in SA and how this will help automatically to identify whether a sentiment is positive, negative or neutral.

The terms SA and OM are not restricted to computer science and information technology but other studies like management and social sciences use them [5]

Arab researchers, in Arabic SA field, need a relatively big standard corpus to conduct their studies. There are a number of existing datasets but with limitations such as the small number of reviews or the topics they contain are restricted to Modern Standard Arabic (MSA) and more of these datasets are not available publically. Authors of [11] established a standard Arabic sentiment analysis corpus, for free to be used by researchers in this field. This corpus can be considered as a foundation to build larger Arabic corpora. This corpus is built manually and it allows the users to revise, remove and add its contents. The total number of topics and reviews in this corpus are 250 and 1,442, respectively.

In [12] they created a tool called Arabic Opinions Polarity Identification (AOPI) based on benchmark corpora that consists of 3,015 textual Arabic opinions about three domains (Sport, Food and Weather) collected from Facebook using manually constructed ten Arabic lexicons. They conclude that their tool AOPI by comparison using mentioned dataset is more effective than other two free SA tools (SocialMention and SentiStrength).In [13]they provided an empirical analysis and comparison of the most recent classifiers for social media such as Naïve Bayes, Bagging, Random Forest, Decision Tree, Support Vector Machine and Maximum entropy for analyzing the good performance and better information retrieval. They designed a comprehensive comparative framework and used benchmark datasets (UCI, KEEL) for comparison purpose. When they analyzed the result, they found that SVM[6]performs much better as compared to other. There are two SA popular approaches; machine learning SA (MLSA) and lexicon-based SA (LBSA).

Machine learning sentiment analysis (MLSA)

MLSA is intended for supervised learning approaches that need pre-annotated text data set in order to train and test some selected features. In this approach, we usually find a model using classification or clustering technique, depending on training data, and then evaluate the model using test data set.

Lexicon based sentiment analysis (LBSA)

LBSA is the process of detecting text implied opinions, using sentiment lexicon, which is, usually, a list of three types of terms, positive terms, negative terms, and negations. Negations are sort of stop-words that don't bear sentiment meaning, but they affect sentiment terms, which may lead to inverse polarity context. The simplest example of negation is 'No' word, in English or its equivalent in Arabic, also any contradictory terms such as 'but', 'non'…etc. [14].

2.2.3 Feature Selection Methods

Selecting a subset of optimal features process is called as feature selection. As mentioned earlier there are two machine learning approaches (supervised and unsupervised) class label is used in supervised machine learning approaches, but in unsupervised machine learning approaches a feature selection is more difficult since class labels are not present.

Feature selection methods are normally used to reduce the dimensionality of datasets to improve the performance of the classification, or to reduce the processing time, or both[15].In this work, the authors present a feature selection algorithm, based on terminology, extracted from the statistics of compound words, to reduce the high dimensionality of the feature space. They compare their algorithm with six well-known feature selection methods including Information Gain, Chi-Square, Gini Index, Support Vector Machine-Based, Principal Components Analysis and Symmetric Uncertainty using three Arabic standard datasets and with three classification algorithms. When they come to the result, they found that their proposed method behaves higher performance in terms of accuracy than the considered approaches with a 6-10% gain in the macro-average.

In [16] they propose a hybrid approach for sentiment analysis based on deep learning with features weighting for Arabic tweets. In this work, to select the most significant features, the features weighting methods are used in the pre-processing stage and the deep leaning is employed to analyze the sentiment of Arabic tweets based on the selected features. The performance of proposed approach based on deep learning with Chai-square method is compared with the performance of some other

classifiers SVM, (decision trees) DT and (Neural Networks) NN. When the result is found, the authors conclude, that their proposed hybrid approach outperforms other classifiers and achieved the highest performance in terms of accuracy and precision of 90% and 93.7% respectively.

Authors of [17] present a wrapper based unsupervised feature selection method with the modified binary bat approach with k-means clustering algorithm. They introducea mutation operator in their algorithm and they conclude that their proposed algorithm can identify a minimal number of selected features with high accuracy when they compare it with some other methods as SVM, NB and DT classifiers.

Opinion mining basics are based on text mining and information retrieval fundamentals, which begin with constructing a vector space model (VSM)[15], it is an *n* by *m* matrix, where we have *n* documents consisted of *m* unique terms in the whole data set; i.e. a corpus, as a result of this, each document is represented in a vector, from rows vectors, also each unique term is represented in a vector from columns vectors. For representing VSM, we had some options and there are two popular methods: Term-Document Matrix (TDMX) and Term-Frequent Matrix (TFMX)[6] that we will call them in the rest of this paper as traditional feature selection methods:

tdmx is a binary VSM matrix in which each cell indicates if the term exists on the corresponding document or not, if the term exists, the cell takes the value of *one*, else it takes *zero*.

tfmx is similar to TDMX; however, TFMX values are integers not binary. In TFMX each cell is zero if the term does existed in the corresponding document, otherwise the cell takes the number of times often occurs in that document. There are further methods for representing documents and terms in a corpus, although, we are contented with the previous traditional methods because many experiments utilized them such as the most popular methods in this context, rather than other methods. In [6] they conclude that the mentioned pervious methods are exceeded another traditional techniques in the area of opinion mining.

3. PROPOSED APPROACH

Proposed approach is based on feature selection mechanism. It concentrates on utilizing opinion properties in each text document, then extracting a term weight for each unique term in the whole corpus.

Each corpus includes a number of text documents, each document consists of a number of terms, since we determine a weight that is related to each term, and this weight is computed incrementally, depending on list of terms which are an opinion lexicon.

Unique terms in the whole corpus mostly exist in number of documents, each term has a different distribution on these documents, so we make a weight for each term, this weight is used then with the traditional methods for enhancing the accuracy of opinion mining discovery, which reflects a successful result as illustrated in the results section.

Opinion's word weight (oww)

Our opinion lexicon consists of two lists of terms, negative terms list and positive terms list, both lists are used for finding out document polarity, as it is explained in LBOM approach.

Two matrices are generated using the lexicon two lists**ptfmx** and **ntfmx**. Both matrices are generated form **tfmx**matrix, so each one of them is*n* by *m* matrix, where *n* is number of documents and *m* is the number of unique terms. Positive **tfmx**is called **ptfmx** because it is produced by matching positive term list with unique document terms, and then it is found out if the term exists, we then match the existing terms with each document, only the term existing in the document keeps its **tfmx** value, each other terms values are converted to zero. The same is done to generate **ntfmx** that is the TFMX of negative terms list. We use**ptfmx** and **ntfmx**to produce Opinion's word weightFor the purpose of computing OWW, we begin by computing positive document polarity using positive termslist, and of course negative terms list, as using the formula in the following equations:

$$Pdoc(i) = \sum_{j=0}^{n} (ptfmx(i,j)) \quad (7)$$

$$\mathrm{Ndoc(i)} = \sum_{\mathrm{j=0}}^{\mathrm{n}} (\mathrm{ntfmx(i,j)}) \quad (8)$$

Where **Pdoc(i)**, **Ndoc(i)**are positive and negative polarity of document (*i*), **ptfmx** and **ntfmx**are positive **tfmx** and negative **tfmx** matrices respectively and ***n*** is the number of documents in the corpus.

After that, OWW vector of allunique terms is computed by:

$$\mathrm{oww(j)} = \sum_{\mathrm{j=0}}^{\mathrm{m}} \mathrm{Ndoc(i)} * \mathrm{tdmx(i,j)} - \sum_{\mathrm{j=0}}^{\mathrm{m}} \mathrm{Pdoc(i)} * \mathrm{tdmx(i,j)} \quad (9)$$

Where ,***j*** is a term of *m* unique dataset terms. In this equation we generate a value related to each term in the corpus, depending on its distribution in corpus documents and the polarity of each document.

Applying equation (3) to each unique term in the corpus is generating **oww** vector in *m* length, for using this vector as a feature selection represents corpus documents, we combine it with both traditional OM feature selection techniques, this leads to generate a new feature selection approach which achieve enhanced accuracy displayed and itis discussed in the results section.

Word's weights with TDMX

We have produced **owtd** *an m* by *n* matrix; since we multiplied each columns vector in **tdmx** matrix as ***n*** by ***m*** matrix by OWW vectorby dot multiplying, so each element in **owtd**is generated by the following equation:

$$owtd(i,j) = TDMX(i,j) * oww(j) \quad (10)$$

Where ,***i*** indicates to document (*i*) and ***j*** indicates to term (*j*) in the corpus.

Word's weights with TFMX

The matrix **owfd**is generated likewise **owtd** using the following equation:

$$owfd(i,j) = TFMX(i,j) * oww(j) \quad (11)$$

So, as it is illustrated in the next section the results are based on **owtd** and **owfd** as hybrid of new feature selection approach based on **oww** vector and traditional methods.

4. EXPERIMENTS AND RESULTS

Proposed approach is experimented, as using the two new feature selection approaches: **owtd** and **owfd**. Then, we compare the new approaches with two traditional methods; i.e. **tdmx** and **tfmx**. Each feature selection are extracted from one of three corpora, two are generated from LABR corpus, large Arabic book review data set, and the last is generated from OCA; open Arabic corpus, which is a movie review data set, both datasets are written in Modern Arabic Standard (MAS). Although we take all OCA dataset, we cannot use all LABR due to its huge amount, therefore we extract two different corpora from LABR; balanced and non-balanced ones, balancing is based on the number of positive comparing to negative documents in each corpus.

Moreover, we have experimented with a proposed approach using four machine learning techniques, the results of our experiments are listed in the following subsections, and the discussion and graph illustration of the results are shown in the next section.

4.1 Supervised learning techniques results:

We have used two supervised classifiers; NBC and KNN, and two experiments are conducted on the three corpora, using the four feature selection methods. The outcome has given the results in tables (1) and (2). It is obvious that our proposed approaches outperform the traditional feature selection methods, in term of testing accuracy percentages.

Table(1) NBC Accuracy on the three corpora

	TDMX	TFMX	owtd	owfd
unBSlabr	73.00 %	71.30 %	74.78 %	74.14 %
blnSlabr	73.22 %	74.00 %	75.00 %	68.00 %
OCA	79.65 %	81.76 %	**83.11 %**	81.17 %

Table(2)KNN accuracy on the three corpora

	TDMX	TFMX	owtd	Owfd
unbSlabr	64.88 %	65.36 %	66.32 %	65.28 %
blnSlabr	62.75 %	60.50 %	64.75 %	64.00 %
OCA	73.00 %	80.10 %	**83.60 %**	81.20 %

4.2 Results for Unsupervised learning techniques:

Two clustering techniques are utilized, as non-supervised learning techniques; k-means and fuzzy c-means, they have given the results in tables (3,4).It is obvious that new approaches outperform the results of traditional ones, as well as in the supervised learning.

Table (3)K-MEANS accuracy on the three corpora

	TDMX	TFMX	Owtd	owfd
unbSlabr	52.80 %	55.20 %	62.04 %	61.44 %
blnSlabr	50.15 %	50.85 %	52.05 %	51.10 %
OCA	59.00 %	56.00 %	67.80 %	56.40 %

Table (4) C-MEANS accuracy on the three corpora

	TDMX	TFMX	Owtd	owfd
unbSlabr	55.64 %	56.12 %	56.12 %	56.12 %
blnSlabr	53.75 %	52.45 %	54.90 %	54.00 %
OCA	59.40 %	60.00 %	63.0	58.40

5. RESULTS DISCUSSIONS

Accuracy measure is used for evaluating ML training phase; it is a very popular measure for estimating the success of models we have built. Accuracy measure evaluate the ratio of success found in patterns compared to all discovered patterns via test data set, which point to the strength of the constructed model for each learning operation. We have considered each accuracy percentage bigger than or equal to 50 percentage as a successful learning, because the main purpose is to compare feature selection mechanisms, for evaluating proposed feature selection approach in Arabic language text data. However, this goal is achieved as showed in the results tables 1,2,3, and 4, the Fig1 is made better illustration of the advanced accuracy of **owtd** and **owfd** feature selection comparing to traditional methods, this is happening in all experiments, regardless used technique, and the text data set.

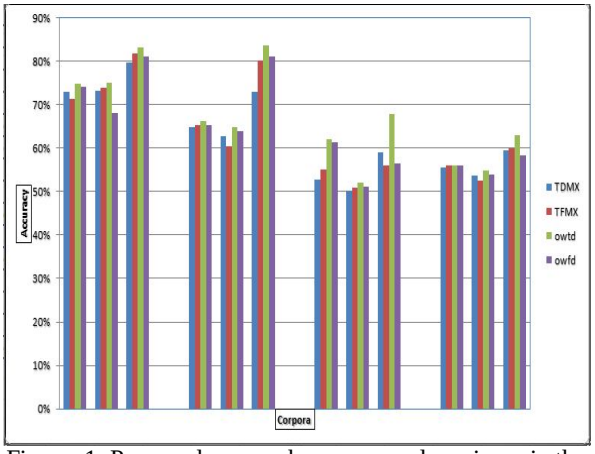

Figure 1: Proposed approach accuracy advancing, via the four experiments.

NBC have gained the best accuracies on each corpus, looking at Figs 2, 3, and 4. Although, it faced a little competition by KNN, in the case of OCA corpus, as illustrated on Fig 4, it still outperforms all the techniques. It is used to know that NBC is a better classifier, especially in text mining, where it based on Probability function, it seems consistent with the bag-of-word; the feature selection construction principal, also as in equations (1) and (2), they are customized for text mining purpose, which must help in increasing NBC results.

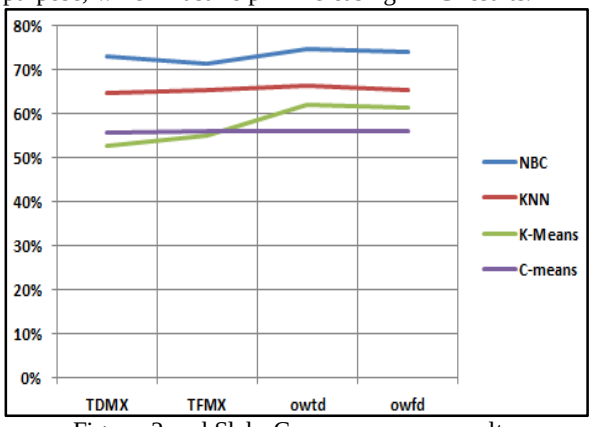

Figure 2: unbSlabr Corpus accuracy results.

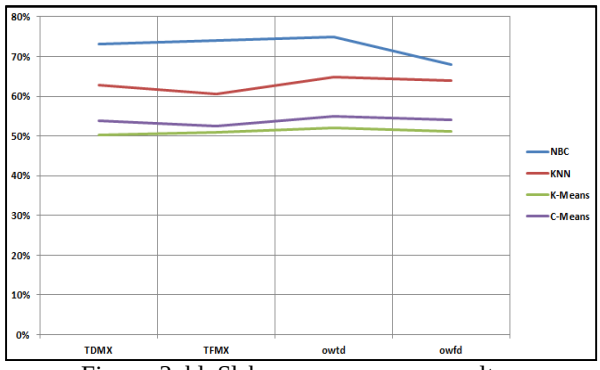

Figure 3: blnSlabrcorpus accuracy results.

On the other hand, accuracy results are varied per corpus, this is obvious comparing Figs 2 and 3 to Fig 4, since Fig 2 illustrated ***unbSlabr***; unbalanced corpus derived from LABR, and Fig 3 show the ***blnSlabr*** corpus, the balanced one, in term of distributing number of negative and positive classes over the corpus, whereas Fig 4 illustrate another corpus; OCA open Arabic corpus, notice the increasing of accuracy in latter corpus.

OCA corpus is an Arabic language movie reviews text data set, it has a long-sized text records, despite of the less size; only 500 record. However, both ***blnSlabr***and ***unbSlabr***are extracted from book review website, they are bigger than OCA in number of records, but they have small-sized text records, which might affect the results of finding sentiment patterns. The latter two corpora are from the same domain, so their results are closed, although, they are less accuracy than the corpus of movie reviews domain, because OCA are more specialized than book review writers. OCA writer has used sophisticated terms, LABR derived two corpora writers are just book reader, and they gave impressions of the reading books, in fewer terms and simple language, comparing to movie review texts.

The persistent point in Figs 2, 3, and 4 is the area of our proposed approach increasing; it is obvious over the Figs the increasing accuracy of the **owtd** feature selection, which asserts the basic goal of our work, a new feature selection approach over all corpora over all techniques.

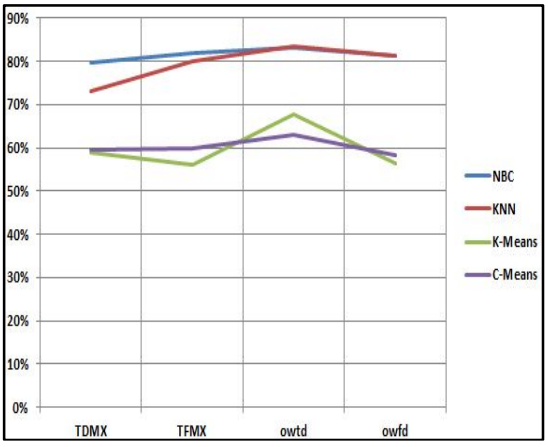

Figure 4: OCA corpus accuracy results.

6. CONCLUSION

We had implemented a number of experiments via four machine learning techniques, on the same three Arabic language corpora. We have aimed to test our rule-based feature selection approach. The four ML techniques are NBC, and KNN; as supervised learning, and k-means, and fuzzy c-means; as unsupervised learning techniques. We have carried out four experiments on three corpora, each experiment has tested four selection method, on each corpora using each technique, the results of all experiments have quiet consistency, as well as pointing to the outperforming role of the proposed opinion mining approach on ML feature selection, over traditional feature selection approaches. The results discussion shows the proposed approach advantage, presents the variety of techniques results, and discusses the aspects of each corpus domain, which draws its effect on the results and the graphs Figs that illustrate these results.

REFERENCES

[1] Bing Liu. **Sentiment Analysis and Opinion Mining**, Morgan & Claypool Publishers, May 2012.

[2] Nasukawa T. and Yi J., **Sentiment Analysis: Capturing Favorability using Natural Language Processing**, Proceedings of the 2nd International Conference on Knowledge Capture, Sanibel Island, 2003, 70-77 pp

[3] Alqasemi Fahd, Amira Abdelwahab, Hatem Ahmed Abdelkader, **Adapting Domain-specific Sentiment Lexicon using New NLP-based Method in Arabic Language**, International Journal of Computer Systems, Vol. 03, Issue 03, March 2016, 188-193 pp

[4] Buket ERŞAHİN, Özlem AKTAŞ, Deniz KILINÇ, Mustafa ERŞAHİN, **A hybrid sentiment analysis method for Turkish**, Turkish Journal of Electrical Engineering & Computer Sciences, 2019.

[5] Al-Ayyoub M., Bani-Essa S., and Alsmadi I., "**Lexicon-Based Sentiment Analysis of Arabic Tweets,"** International Journal of Social Network Mining, vol. 2, no. 2, 2015, 101-114 pp

[6] Rushdi-Saleh M, Martin-Valdivia M, **Urea-Lopez L, et al. OCA: Opinion corpus for Arabic.** Journal of the American Society for Information Science and Technology. 2011;10:2045-2054.

[7] Nabil M, Aly M, Atiya A. LABR: **A large scale Arabic book reviews dataset.**CoRR. 2014; abs/1411.6718.

[8] TheodoridisSergios, **Pattern recognition**, 4th edition, 2009.

[9] Fahd A. AlqasemiȦ, Amira AbdelwahabḂ and Hatem Abdelkader, **An Image Retrieval Approach based on Colored Regions Features And a Clustering Technique**, International Journal of Computer Systems (ISSN: 2394-1065),

Volume 03– Issue 03, March, 2016, Available at http://www.ijcsonline.com.

[10] Ravi Sanakal and Smt. T Jayakumari, **Prognosis of Diabetes Using Data mining Approach-Fuzzy C Means Clustering and Support Vector Machine**, International Journal of Computer Trends and Technology, Vol. 11, No 2, May 2014, 94-98 pp

[11] Mohammed Al-Kabi, Mahmoud Al-Ayyoub, Izzat Alsmadi and Heider Wahsheh, **A Prototype for a Standard Arabic Sentiment Analysis Corpus**, International Arab Journal of Information Technology, Vol. 13, No. 1A, 2016, 163- 170 pp

[12] Mohammed Al-Kabi, IzzatAlsmadi, RawanKhasawneh, and HeiderWahsheh, **Evaluating social context in Arabic opinion mining**, International Arab Journal of Information Technology, Vol. 15, No. 6, November 2018, 974-983 pp

[13] Maria Hameed, Faizan Tahir, M. Ali Shahzad, **Empirical comparison of sentiment analysis techniques for social media**, International Journal of Advanced and Applied Sciences, 5(4), 2018, 115-123 pp

[14] Taboada, M., BrookeJ., Tofiloski M., Voll, K., Stede M., **Lexicon-based methods for sentiment analysis**, Computational linguistics, 37(2), , 2011, 267-307 pp

[15] Aisha Adel, Nazlia Omar, Mohammed Albared , and Adel Al-Shabi, **Feature Selection Method Based On Statistics of Compound Words for Arabic Text Classification**, International Arab Journal of Information Technology, Vol. 16, No. 2, March 2019, 178-185 pp

[16] Altyeb Altaher, Hybrid **approach for sentiment analysis of Arabic tweets based on deep learning model and features weighting**, International Journal of Advanced and Applied Sciences, Vol. 4, no 8 2017, 43-49 pp

[17] Rajalaxmi Ramasamy and Sylvia Rani, **Modified Bat algorithm for feature selection in unsupervised learning**, The International Arab Journal of Information Technology, Vol. 15, No. 6, November 2018, 1060-1067pp.