Journal of Computer Science

Original Research Paper

Arabic Sentiment Classification using MLP Network Hybrid with Naive Bayes Algorithm

[1]Mohammad Subhi Al-Batah, [2]Shakir Mrayyen, and [3]Malek Alzaqebah

[1,2]Faculty of Science and Information Technology, Jadara University, Irbid, Jordan
[3]Faculty of Science, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia

Article history
Received: 17-01-2018
Revised: 02-04-2018
Accepted: 07-08-2018

Corresponding Author:
Mohammad Subhi Al-Batah
Faculty of Science and Information Technology,
Jadara University, Irbid, Jordan
Email: dralbatah@gmail.com
albatah@jadara.edu.jo

Abstract: Sentiment analysis has recently become one of the growing areas of research related to text mining and natural language processing. Sentiment analysis techniques are increasingly exploited to categorize the opinion text to one or more predefined sentiment classes for the creation and automated maintenance of review-aggregation websites. Most of the current studies related to this topic focus mainly on English texts with very limited resources available for other languages like Arabic. The complexities of Arabic language in morphology, orthography and dialects makes sentiment analysis for Arabic more challenging. In this study, the Naive Bayes algorithm (NB) and Multilayer Perceptron (MLP) network are combined with hybrid system called NB-MLP for Arabic sentiment classification. Five datasets were tested; attraction, hotel, movie, product, and restaurant. The datasets are then classified into positive or negative polarities of sentiment using both standard and combined system. The 10-fold cross validation was employed for splitting the dataset. Over the whole set of experimental data, the results show that the combined system can achieve high classification accuracy and has promising potential application in the Arabic sentiment analysis and opinion mining.

Keywords: Big Data, Social Networks, Machine Learning, Sentiment Analysis, Arabic Language Classification, Naive Bayes Algorithm, Multilayer Perceptron Network

Introduction

Sentiment analysis encompasses the vast field of effective classification of user generated text under defined polarities. There are several tools and algorithms available to perform sentiment detection and analysis including supervised machine learning algorithms that perform classification on the target corpus, after getting trained with training data. Lexical techniques which performs classification on the basis of dictionary based annotated corpus and Hybrid tools which are combination of machine learning and lexicon based algorithms (Duwairi, 2015). There has been a constant rise in the use of many social networks, such as TripAdvisor, Yelp, Foursquare, Booking, and Twitter. In such networks, users can write their opinions about services, food, places to visit, hotels, etc.

The fields of text mining and information retrieval for the Arabic language has been the interest of many researchers, and various studies have been carried in these fields resulting in diverse resources, corpora, and tools available for implementing applications.

Abdul-Mageed *et al*. (2012), have proposed a system called SAMAR for Subjectivity and Sentiment Analysis (SSA), which requires identifying whether the text is objective or subjective before identifying its polarity. The proposed system uses the SVM^{light} algorithm for classification and the dataset they used was collected from four different genres of social media websites: chat, Twitter, Web forums and Wikipedia Talk Pages. Their experiments showed how difficult and complex the characteristics of Arabic language in SSA.

Alhazmi and Salim (2015) introduced a supervised approach to extract the opinion target from Arabic Tweets. To build a training dataset, they manually tagged the opinion target in 500 collected Tweets. After pre-processing Tweets, each word was considered as a training vector defined by POS, named entities, English words and tweet hash tags features. Classification was carried out by specifying that a given word is either an

opinion target or not. Experiments were undertaken using three classifiers: Naïve Bayes, Support Vector Machine and K-Nearest Neighbour. The best result was reached using the K-Nearest Neighbour classifier with an F-Measure of 91%.

In Abu Hammad and El-Halees (2015) proposed a supervised approach for detecting opinion spams in Arabic. This approach combines techniques from data and text mining. The authors collected 2848 Arabic reviews from online accommodation booking websites namely, booking.com, tripadvisor.com, and agoda.ae. In addition, they integrated their dataset into a coherent form data and labelled each review with a spam or a non-spam label. For classification purposes, they used NB, K-NN and SVM classifiers with 10 folds cross-validation. Although their system was limited to hotel reviews, it was able to generate a high accuracy by combining data and text classification.

An Arabic dataset consisting of 500 movie reviews was built by Rushdi-Saleh *et al.* (2011). The authors used SVM and NB in their study. They started their study by preprocessing the collected dataset. The conducted preprocessing operations included manual spelling correction, stop-words removal, stemming, and N-Gram tokenization. Although their experimentation results showed accuracy close to 89%, the size of the dataset they used was small compared to other datasets used in other English-based studies.

Al-Subaihin *et al.* (2014) have created and implemented a lexicon-based sentiment analysis tool for colloquial Arabic text. They applied it on a dataset comprised of Arabic forums comments and newspaper articles written in Arabic.

Nabil *et al.* (2015) presented a 4 way sentiment classification that classifies texts in four classes: objective, subjective negative, subjective positive and subjective mixed. Their dataset has 10,006 Arabic Tweets manually annotated using Amazon Mechanical Turk (AMT) service. They applied a wide range of machine learning algorithms (SVM, MBN, BNB, KNN, stochastic gradient descent) on the balanced and unbalanced datasets. However, using n-grams as unique features in multi-way classification did not give good results.

A lexicon-based approach was proposed in Mourad and Darwish (2013) to perform subjectivity classification of both MSA news articles and dialectal Arabic microblogs from Twitter. In order to build a large lexicon, the authors use two available lexicons: MPQA which is an existing English subjectivity lexicon, and ArSenti, a manually created Arabic lexicon. The first one is translated into Arabic using Machine Translation, and the second is automatically extended using a random graph walk method. All the words in Tweets and in the lexicon were tokenized and stemmed. Polarity stems, as indicated in the lexicon, were used as input feature vector to the learning module.

Mountassir *et al.* (2012) conducted a binary sentiment classification using three classifiers: NB, SVM and KNN. Two corpora were used: the first is developed by these authors and is composed of two domain-specific datasets (movies and sports). The second is OCA, a corpus of movie reviews developed by Rushdi-Saleh *et al.* (2011). Before the classification phase, the authors performed a pre-processing task by removing stop words, separating words from their clitics, eliminating terms used only once or twice in the dataset, and by replacing words by their stems. The authors found out that pre-processing, n-grams combination, and presence-based weighting improve the classification performance.

On the other hand, Aly and Atiya (2013) created a Large-scale Arabic Book Review (LABR), a dataset of over 63,257 book reviews collected from www.goodreads.com. Reviews with rating 4 or 5 were labelled as positive. Negative reviews were those with rating 1 or 2 while reviews that were rated 3 were considered neutral. Since the number of positive reviews (42,832) was much larger than that of negative reviews (8224), they applied machine learning in both balanced and unbalanced data using SVM, MNB and BNB as algorithms and n-grams as features. For sentiment polarity classification, the evaluation of their dataset achieved quite good results (~90% accuracy), but for rating classification there is much room for improvement (~50% accuracy).

For their parts, El-Beltagy and Ali (2013) proposed a lexicon-based approach to establish a sentiment classification of Egyptian Arabic texts. After building a lexicon of 4392 terms, the authors used two datasets (Twitter dataset of 500 tweets and Dostour dataset of 100 web comments) to evaluate two unsupervised classification algorithms. The first calculates one score for each document by adding up weights of negative and positive terms. The second algorithm assigns a positive and a negative weight to each term in the lexicon and calculates positive and negative scores for each document. The authors achieved good results using the two algorithms on a Twitter dataset (83.8% accuracy).

Shoukry and Refae in (2012) worked on a tweet dataset composed of 1000 tweets (500 are positives and 500 are negative). They dealt with sentence-level sentiment analysis since tweets length is restricted to 140 characters. Though their work lacks handling the neutral cases and exploits a small corpus, they explored the direction of Arabic dialects and appended some words from the Egyptian dialect alongside the MSA ones. For the preprocessing phase, they applied Unigram-based and Bigram-based features extraction techniques and concluded that there is no difference in the results. The approach followed in this paper was corpus-based

(supervised approach), where SVM and NB were used for polarity classification. The results showed that SVM outperformed NB in sentiment analysis with an accuracy of 72.6% regardless to the feature extraction technique used (whether it is Unigram-based or Bigram-based).

In El-Makky *et al.* (2015) combined Sentiment Orientation (SO) algorithms with a machine learning classifier to propose a hybrid approach. For each document in a Twitter dataset, they used the lexicon-based approach to compute Sentiment Orientation scores. These scores were integrated with different features such as unigrams, language independent features, Tweets-specific features and stem polarity features so as to create an input feature vector for the SVM classifier. This combination of the Machine Learning classification approach and the lexicon based approach led to slightly better results than a one-approach result (accuracy 84%).

Al-Smadi et al. (2015) set up a benchmark dataset of Arabic reviews. Their Human Annotated Arabic Dataset (HAAD) contains 1513 reviews selected from LABR (Aly and Atiya, 2013) and manually annotated. Following SemEval2014 guidelines proposed in Pontiki *et al* (2014), annotators were asked, in the first phase, to identify aspect terms and provide the polarity for each aspect term. In the second phase, annotators recognized aspect categories and their polarities. As a benchmark dataset, the authors conducted an evaluation baseline of four tasks: aspect term extraction, aspect term polarity, aspect category extraction, and aspect category polarity. The adopted approach uses a majority baseline that assigns the most repeated polarity in the training data to all aspect terms and categories.

Abdul-Mageed and Diab (2014) presented a large scale multi-genre sentiment lexicon. This lexicon is made up of 224,564 entries covering MSA and multiple Arabic dialects. The authors collected and manually tagged two word lists from both Penn Arabic Treebank and Yahoo Maktoob. Lists were automatically developed using Google's translation API of three existing English lexica: SentiWordNet, YouTube Lexicon, and General Inquirer. To expand the lexicon's coverage, they used a statistical method based on PMI to extract other polarized tokens from both Twitter and chat datasets. Despite the large size of the resulting resource, many of the entries are neither lemmatized nor diacritized, which limits the usability of their lexicon.

In their attempt to build Arabic multi–domain resources for Sentiment Analysis, ElSahar and El-Beltagy (2015) proposed a semi-supervised approach to generate multi-domain lexica out of four multi–domains reviews datasets. This method makes use of the feature selection capabilities of SVM to select the most efficient unigram and bigram features. Although the created lexicon covers a variety of domains, it was extracted only from reviews, which restricts its usefulness just for social media Sentiment Analysis.

Badaro *et al.* (2014) set up ArSenL, a lexicon for Arabic sentiments using two approaches based on English SentiWordNet (ESWN). The first method links each term in ArabicWordNet, on the one hand, with ESWN to get sentiment scores, and on the other hand with SAMA (Standard Arabic Morphological Analyser) to find the correct lemma forms. In the second approach, English glosses associated with SAMA's entry were explored automatically to find the most similar synset in ESWN. The union of the two resulting lexica has a good coverage but is limited to MSA.

In addition to these corpora, an Arabic Twitter corpus was collected in Refaee and Rieser (2014) using Twitter API and cleaned in a pre-processing phase. Two native speakers of Arabic annotated manually 8868 Tweets using four labels: neutral, mixed, positive and negative. Morphological, syntactic, and semantic features were also added to the annotation.

In Elarnaoty and AbdelRahman (2012), the authors explored the problem in Arabic news articles using three different approaches. The first approach is semi-supervised and uses a set of handcrafted patterns. POS tags and key phrases were used to define 43 patterns, which were chosen to run a pattern matcher code and identify opinion sources on the tested data. The second is a supervised machine learning approach that uses the Conditional Random Field classifier (CRF). To train this classifier, the authors used features such as surrounding words, POS, Named Entity and sentiment words. The third approach is a combination of the two previous approaches using patterns such as CRF features. Their experiments showed that the CRF outperforms patterns in terms of recall and precision. Moreover, adding patterns as a feature to CRF is insignificant compared to other features such as the Named Entity feature.

A preliminary study that was concerned with opinion spam detection was conducted by Wahsheh *et al.* in (2013). The study was based on 3090 Arabic opinions collected from Yahoo-Maktoob social network. The authors employed ACLWSDS, which is an Arabic spam URL detection system developed in Wahsheh *et al.* (2012). Opinions containing a URL were classified, either as high-level spam if the URL was considered as spam by ACLWSDS, or as low-level spam if the URL was considered as non-spam. In the absence of URLs and some specific metrics, the opinion was categorized as a non-spam. Evaluating this method with SVM algorithm achieved favourable results. However, using only the URL filtering technique may not be efficient.

On the other hand, Naive Bayes (NB) classifier is a probabilistic classifier that can be categorized as supervised classification method. NB applies Bayes' theorem with strong independence assumptions, which

showed to be effective, simple, fast and high accuracy in text classification (Duwairi and Qarqaz, 2014). Multilayer Perceptron (MLP) network is one of the most popular and important networks of Artificial Neural Network (ANN). MLP has a strong associative memory and prediction capability after training. To increase the performance classification for the Arabic sentiment analysis datasets, this study aims to combine the NB classifier with MLP network.

This paper consists five other sections that organized as follows: next section describes the datasets used in Arabic sentiment analysis, followed by discussion about the Naive Bayes (NB) classifier, Multilayer Perceptron (MLP) Network, and the proposed hybrid NB-MLP. Then the performance evaluation and the results finding are reported, finally, the conclusion and recommendations are presented.

Sentiment Analysis and Datasets

Sentiment analysis is a current research area in text mining. It is the stem of natural language processing or machine learning methods. It is the important sources of decision making and can be extracted, identified, evaluated from the online sentiments reviews or tweets.

Reviews datasets covers five domains are considered in our experiment as follow: (ElSahar and El-Beltagy, 2015; Nabil *et al*., 2014; Pang *et al*., 2002).

- **Attraction (ATT):** Dataset of Attraction Reviews scrapped from TripAdvisor.com, which contains 2K Arabic reviews
- **Hotel Reviews** (**HTL)**: For the hotels domain, the reviews were collected from the TripAdvisor website, which contains 15K Arabic reviews, were written by 13K users for 8100 Hotels
- **Restaurant Reviews (RES):** For the restaurants domain, the reviews were scrapped from Qaym 8.6K Arabic reviews were collected
- **Movie Reviews (MOV):** For the movies domain, the dataset contains 1.5K reviews were collected from Elcinemas.com which cover around 1K movies
- **Product Reviews (PROD):** For the Products domain, a dataset of 15K reviews were written by 7.5K users and cover 6.5K products, the dataset were collected from Souq.com. The reviews are from Egypt, Saudi Arabia and the United Arab Emirates

The description of the datasets properties, (such as total number of reviews or tweets, number of positive and negative reviews) are presented in Table 1. As shown in Table 1, we have 2154, 13420, 1353, 3962, and 8522 reviews for attraction, hotel, movie, product, and restaurant, respectively. Each dataset contains positive and negative reviews. As example, the attraction dataset contains 2154 text files in which 2073 labelled as positive reviews and the rest 81 labelled as negative reviews.

Table 1: Sentiment Analysis Dataset

Dataset	Total No. of reviews	Pos. reviews	Neg. reviews
ATT	2154	2073	81
Hotel	13420	10773	2647
Movies	1353	969	384
Products	3962	3100	862
Res	8522	5938	2416

The Proposed System

A. Naive Bayes Classifier

The Naive Bayes (NB) classifier is commonly used for the review classification. The algorithm can determine the rear possibilities of the classes to relate the review with the help of a feature vector table. The review is then assigned to the class with the maximum rear possibility. Normally employed two models of the Naïve Bayes approach, multinomial and Bernoulli's multivariate, for text classification. The Naïve Bayes is a stochastic model for generating documents, which follows the Bayes' rule as follow (Leung, 2018):

$$P\left(c_i \left| d_j \right.\right) = \frac{P\left(c_i\right) * P\left(d_j \left| c_i \right.\right)}{P\left(d_j\right)} \quad (1)$$

Where:

$c_{i:}$	Specific class that might be positive or negative
d_j:	Document (text) to be classified
$P(c_i)$ and $P(d_j)$:	Prior probabilities
$P(c_i \mid d_j)$ and $P(d_j \mid c_i)$:	Posterior probabilities

In this work we use the multinomial model. Since the multinomial model is better than the multivariate Bernoulli model in text classification (Shimodaira, 2018).

NB multinomial model responsible for capturing the word frequency in documents (Radovanovic and Ivanovic, 2018). The Maximum Likelihood Estimate (MLE) is a method of estimating the parameters of given training data based on relative frequency (Dhande and Patnaik, 2014). MLE attempts to find the parameter values that maximize the most likely value (using the

likelihood function). For the prior probability, this estimation is shown in Equation 2:

$$P(c_i) = \frac{Nc}{N} \quad (2)$$

Where:
Nc: The number of documents in class c_i
N: Total number of documents

Multinomial model assumes that all attributes are independent of each other given the context of the class. The problem with the MLE is that the zero for the number of words in the class c_i did not occur in the training data. The training data are insufficient to represent the frequency of rare occurrence words. In order to avoid zero probability, add-one smoothing (or Laplace smoothing), this simply adds one to each count. The conditional probability $P(W_k|c_i)$ is simply multinomial distribution presented in Equation 3, which shows the relative frequency of the word *W* in documents belonging to class c_i (including multiple occurrences of a term *W* in that document):

$$P(W_k|c_i) = \frac{n_k(c_i)+1}{n(c_i)+|V|} \quad (3)$$

Where:
$n(c_i)$ = Is the number of words in the class c_i (pos or neg)
$n_k(c_i)$ = The number of times word k occur in class c_i
V = Is the vocabulary which stores the unique words
|V| = The number of unique words

The classification of new text is according to the value of naive bayes classifier V_{NB}, which presented in Equation 4:

$$V_{NB} = \arg\max_{cj \in v} P(c_j) \prod_{w \in words} P(w_i|c_j) \quad (4)$$

where, c_j is one of the class from class c and w_i is i^{th} word.

Example we have target concept whether a document is positive or negative:

1. Each document is represented as vector of words storing all unique words
2. Learning, use training examples to estimate the probability of the positive outcomes negative outcomes and probability of each document with each word with positive or negative as follow:
    - P(pos)
    - P(neg)
    - P(word in doc|pos) positive probability of word
    - P(word in doc|neg) negative probability of word

Table 2 presents an as the example of restaurant review. Using NB classifies restaurant review as positive or negative category. Table 2 shows set of classified sentences as training set.

Table 2: Set of classified sentences

No	Document	Translation	class
1	المطعم جيد	It is a good restaurant	pos
2	الطعام هنا جيد و لذيذ	The food here is good and delicious	pos
3	خدمة سيئة هنا	bad service	neg
4	عمال المطعم متعاونين	The workers here are collaborators	pos

There are 10 unique Arabic words (the word "هنا" [here] occurs twice it counts one word)

First calculate priori probability of pos and neg by using Equation 2:

- p (pos) = 3/4 (three positive out of four)
- p (neg) = 1/4 (three negative out of four)

Second calculate the individual probability of all possible words with positive outcome and negative outcome based on likelihood smoothing NB estimate using Equation 3.

For positive outcome:

P(جيد| pos) = (2+1)/(8+10) = 0.16 (the word occurs twice in doc 1 and 2)
P(المطعم| pos) = (1+1)/(8+10) = 0.11
P(هنا| pos) = (1+1)/(8+10) = 0.11
P(سيئة| pos) = (0+1)/(8+10) = 0.05
P(لذيذ| pos) = (1+1)/(8+10) = 0.11
P(طعام| pos) = (1+1)/(8+10) = 0.11
P(خدمة| pos) = (0+1)/(8+10) = 0.05
:
:
And so on.

For negative outcome:
P(جيد| neg) = (0+1)/(3+10) = 0.07
P(المطعم| neg) = (0+1)/(3+10) = 0.07
P(هنا| neg) = (1+1)/(3+10) = 0.15
P(سيئة| neg) = (1+1)/(3+10) = 0.15
P(لذيذ| neg) = (0+1)/(3+10) = 0.07
P(طعام| neg) = (0+1)/(3+10) = 0.07
P(خدمة| neg) = (1+1)/(3+10) = 0.15
:
:
And so on.

Consider the testing sentence (TS) is:
"المطعم يقدم طعام لذيذ و الخدمة رائعة"

[note that, It's recognized as Positive sentence]

Finally calculate the V_{NB} for the testing sentence (TS) using Equation 4.

	المطعم	يقدم	طعام	لذيذ	الخدمة	رائعة
Pos probability for each word	0.11	-	0.11	0.11	0.05	-
Neg probability for each word	0.07	-	0.07	0.07	0.15	-

p(pos | TS) = 0.11* 0.11* 0.11* 0.05 = 6.66E-05
p(neg | TS) = 0.07* 0.07* 0.07* 0.15 = 5.15E-05

We can observe that the value of p(pos | TS) is greater than the value of p(neg | TS), thus the testing sentence is positive.

B. Multilayer Perceptron (*MLP*) *Network*

One of the popular neural networks for pattern recognition application is the Multilayer Perceptron (MLP) network. The network builds its predictive model using a set of data samples. The MLP network consists of a set of input layer, one or more hidden layers and an output layer as shown in Fig. 1. Each layer contains neurons which are linked to neurons in other layers through the weight and bias values (Al-Batah *et al*., 2015).

The network learns the relationship between pairs of inputs (factors) and output (responses) vectors by altering the weight and bias values. In this work, the MLP network with one hidden layer was used. The rationale for using only one hidden layer stemmed from the fact that the MLP with one hidden layer is sufficiently complete to approximate any continuous function with reasonable accuracy (Cybenko, 1989).

For the one hidden layer of MLP network as shown in Fig. 1, the output of the *j*th hidden neuron at time *t* is given by:

$$u_j^1(t) = F\left(\sum_{i=1}^{n_i} w_{ij}^1 x_i(t) + b_j^1\right); 1 \le j \le n_h \quad (5)$$

where, w_{ij}^1 denotes the weights that connect the input and the hidden layers, b_j^1 denotes the thresholds in hidden nodes and x_i denotes the inputs that are supplied to the input layer. n_i and n_h are the number of input and hidden nodes, respectively. $F(*)$ is the activation function. Various transfer functions such as sigmoid, Gaussian, hyperbolic tangent and hyperbolic secant are used as activation functions in neural networks. The most commonly used one in the perceptron is the sigmoid type, defined as:

$$F(x) = 1/\left(1+e^{-x}\right) \quad (6)$$

where, $F(x)$ is always in the range [-1,1], $\forall x \in R$ (the set of real numbers).

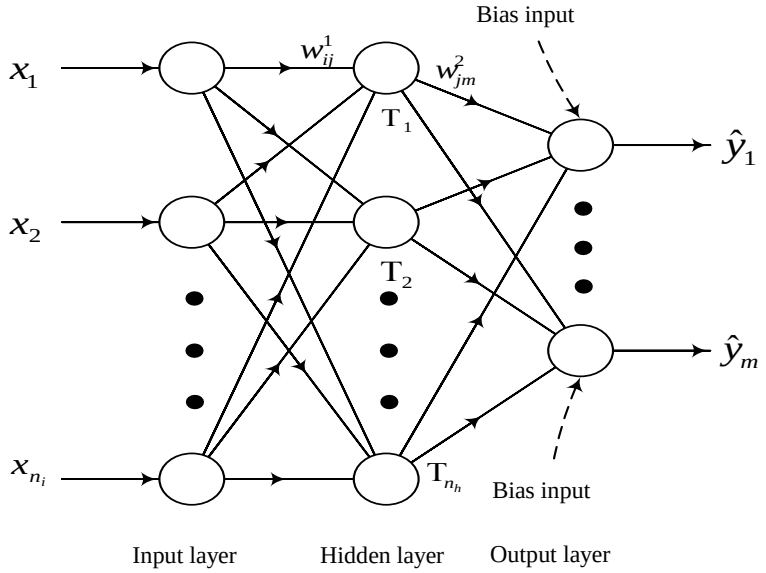

Fig. 1: The MLP network with one hidden layer

The output of the kth neuron, $\tilde{y}_k(t)$ in the output layer is given by:

$$\tilde{y}_k(t) = \sum_{j=1}^{n_h} w_{jk}^2 u_j^1(t); 1 \le k \le m \quad (7)$$

where, w_{jk}^2 denotes the weights of the connections between the hidden and output layers and *m* denotes the number of output nodes.

From Equations (5) and (7), the MLP network with one hidden layer can be expressed by:

$$\tilde{y}_k(t) = \sum_{j=1}^{n_h} w_{jk}^2 F\left(\sum_{i=1}^{n_i} w_{ij}^1 x_i(t) + b_j^1\right); \quad 1 \le j \le n_h, 1 \le k \le m \quad (8)$$

Adjusting the weights between the neurons without a learning algorithm is a difficult task. The backpropagation learning algorithm with momentum was used in this study to reduce the error rate between the actual output and the neural network output results. The algorithm was also used to build up the weight for the input factors (Aggarwal *et al*., 2005). Many researchers used MLP with backpropagation learning algorithms for classification (Baareh *et al*., 2012; Alkhasawneh *et al*., 2014).

C. The Proposed NB-MLP

In this study, the NB is used for classifying the polarity of the documents in five datasets. Some data are complex such as movie reviews dataset. The complexity comes from that the movie reviews are collected from different stories.

The problem, it was not possible for the NB to recognize and categorize the highly complex data with good accuracy. To improve the performance, a combined method for NB and MLP is proposed which is denoted as NB-MLP in this work. In the proposed NB-MLP, the weights of the review datasets and the predicted output of NB is automatically feed to MLP. The MLP re-trains the weight with predicted outputs and enhances the probability of having a correct detection and directly minimizes the classification errors caused by the NB thus refining the weights of the faulty reviews data and this leads to increase the recognition rate. For simplicity, the NB and MLP in the proposed NB-MLP are function together to refine the weights of the input data and improve the performance.

Performance Evaluation and Discussion

Experimental setup contains simulation environment, parameters and performance metrics. Generally, performance metrics are used for calculate one of the metrics like size, execution time, performance accuracy of system (Alkhasawneh *et al*., 2013).

In this study, the performance of the classifiers are evaluated using accuracy parameter. Accuracy is calculated by Equation 9:

$$Accuracy = 100 - \frac{Number\ of\ Incorrect\ Sample * 100}{Total\ Number\ of\ Sample} \quad (9)$$

The proposed approach combined the NB and MLP for Arabic sentiment classification. The proposed system is implemented using C#. Five datasets were tested; attraction, hotel, movie, product, and restaurant.

For each dataset, this study employs a 10 fold cross validation method to arrange the number of the data for training and testing sets (Schaffer, 1993; Al-Batah *et al*., 2010). In this method, the data are partitioned into 10 sized segments or folds. Ten iterations of training and testing are performed. In each iteration, one part of the data is held out for testing while the remaining 9 parts are used for training. The results obtained from the ten runs are then averaged (Al-Batah *et al*., 2009; Isa *et al*., 2008).

In experiment, NB and combined NB-MLP are used for classifying the polarity of the documents in the dataset. In training phase, train the NB classifiers with features for each review dataset. The data review test file is tested using trained classifier. The polarity classification is shown as the result of proposed system. That may be either positive or negative review. Then the weights and the predicted output is feed to the MLP. The MLP with appropriate network structure handles the correlation between input variables.

Table 3 shows accuracy of sentiment analysis using both NB and combined NB-MLP. Based on the results obtained, the proposed NB-MLP produces better overall accuracy compared to NB. It produces 99.8, 85.1, 95.4, 97.3, and 93.1 for attraction, hotel, movie, product and restaurant dataset, respectively. While the NB produces overall accuracy of 96.7, 77.8, 29.3, 90.8, and 85.4 for attraction, hotel, movie, product and restaurant dataset, respectively.

The results in Fig. 2 show that the proposed NB-MLP is able to achieve better classification performance than NB. The NB-MLP outperformed the NB in terms of the percentage of training accuracy by more than 2.5% for attraction, 8.6% for hotel, 64.0% for movie, 3.1% for product and 6.0% for restaurant. In addition, the NB-MLP outperformed the NB with difference of testing accuracy percentage equal to 3.7, 6.0, 68.1, 9.9 and 9.4 for attraction, hotel, movie, product and restaurants,

DOI: 10.3844/jcssp.2018.1104.1114

respectively. The outcomes consistently demonstrate the effectiveness of the proposed NB-MLP for tackling Arabic sentiment classification tasks.

Also, the result shows that the proposed NB-MLP provides better result in complex domain such as movie review dataset. After experiment, the accuracy of sentiment analysis using NB-MLP is 92.6% for training movie review dataset. While, the result of sentiment nalysis using NB is obtained only 28.6% accuracy on training data.

Figure 3 shows result of overall correct samples classification for the five datasets. For movie reviews dataset, NB classifier found 396 correct samples from 1353 reviews. From combing result of NB and MLP, The proposed NB-MLP classifier found 1291 correct samples.

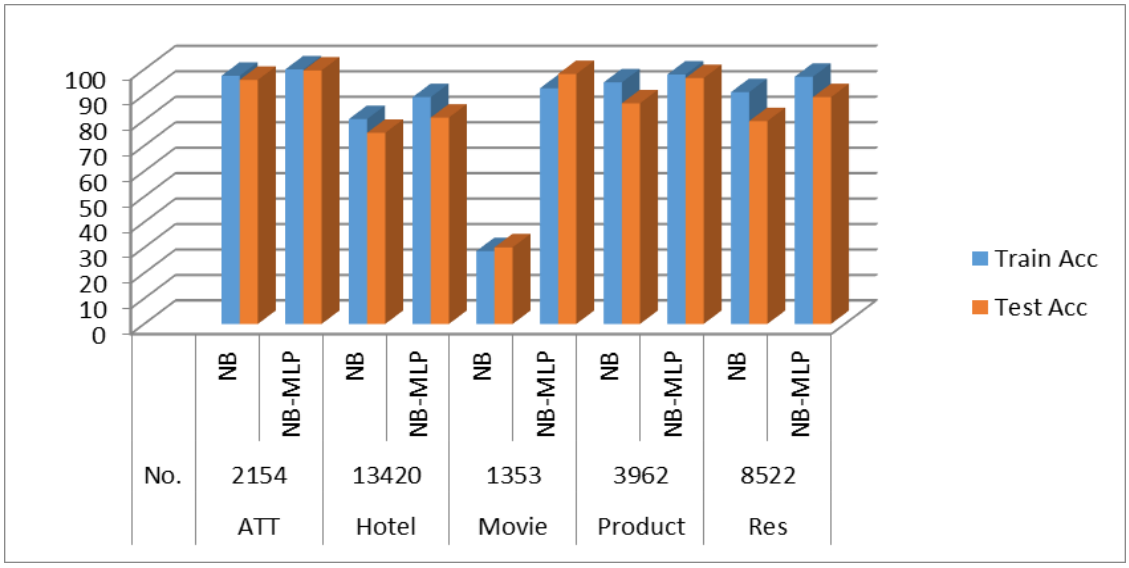

Fig. 2: Train and test accuracy for NB and hybrid NB-MLP

Fig. 3: Overall accuracy for NB and hybrid NB-MLP

Table 3: Result of dataset experiment

Dataset	Total no.	Classifier	Train Acc.	Test Acc.	Overall Acc.	Correct sample
ATT	2154	NB	97.5	95.9	96.7	2083
		NB-MLP	100	99.6	99.8	2150
Hotel	13420	NB	80.5	75.1	77.8	10441
		NB-MLP	89.1	81.1	85.1	11420
Movie	1353	NB	28.6	30.1	29.3	396
		NB-MLP	92.6	98.2	95.4	1291
Product	3962	NB	94.9	86.7	90.8	3597
		NB-MLP	98.0	96.6	97.3	3855
Res	8522	NB	91.1	79.7	85.4	7278
		NB-MLP	97.1	89.1	93.1	7934

As a result, the sentiment analysis using the combined NB-MLP is giving the correct output. The correct output is obtained by combining the predicted output of NB result and MLP classifier result along with actual output using confusion matrix. From assumptions of dependency and independency among features, result of NB is improved by using NB-MLP.

Conclusion

In this study, sentiment analysis of five datasets is conducted; attraction, hotel, movie, product and restaurant. The data are classified into positive or negative polarities of sentiment using NB and Hybrid NB-MLP. Accuracy of sentiment analysis is increased by proposed system from dependence and independence assumptions among features. Over the whole set of experimental data, the performance of the proposed NB-MLP ranked first compared to standard NB with recorded testing accuracy of 99.6, 81.1, 98.2, 96.6 and 89.1 for attraction, hotel, movie, product and restaurant, respectively. Also, the proposed NB-MLP produces better performance than NB with average training accuracy of 100.0% (attraction), 89.1% (hotel), 92.6% (movie), 98.0% (product) and 97.1% (restaurant). These results show the effectiveness of the proposed NB-MLP as compared with NB classifier. In future, apply this work on clustering domain of review dataset for opinion mining applications where the cluster based features are used to address the problem of scarcity of opinion annotated data in a language. Also, the suggestion to use more case studies such as tweets should be done to test the system in order to establish its capability and reliability.

Acknowledgement

This research was supported by Jadara University-Jordan, and Imam Abdulrahman Bin Faisal University-Saudi Arabia. We thank Dr. Mutasem Alkhasawneh, and Dr. Mohammad Klaib for assistance and comments that greatly improved the study and provided insight and expertise that greatly assisted the research.

Author's Contributions

Mohammad Subhi Al-Batah: Proposed the main idea of the data classification based on hybrid NB-MLP. Participated in implementing all experiments, and analyzing the experimental results.

Shakir Mrayyen: Implementing the source code, and analyzing the data using Naive Bayes algorithm. Drawing the results with approprite figures is done by this author as well.

Malek Alzaqebah: Organizing the research study, selecting the approprite datasets, and establishing the state of art. Contributed in writing the article, the simulation results analysis, and interpretation of simulation results.

Ethics

We testify that our research paper submitted to the Journal of Science Publication, title: Investigation of Naive Bayes Combined with Multilayer Perceptron for Arabic Sentiment Analysis and Opinion Mining has not been published in whole or in part elsewhere. All refereces used are cited in this study. There is no ethical issue involved in this article.

References

Abdul-Mageed, M. and M.T. Diab, 2012. AWATIF: a multi-genre corpus for modern standard Arabic subjectivity and sentiment analysis, Presented at the LREC, pp. 3907–3914.

Abdul-Mageed, M. and M.T. Diab, 2014. SANA: a large scale multi-genre, multi-dialect lexicon for Arabic subjectivity and sentiment analysis, Presented at the LREC, pp. 1162–1169.

Abdul-Mageed, M., S. Kübler and M. Diab, 2012. "Samar: A system for subjectivity and sentiment analysis of arabic social media." In Proceedings of the 3rd Workshop in Computational Approaches to

Subjectivity and Sentiment Analysis, pp. 19-28. Association for Computational Linguistics.

Abu Hammad, A. and A. El-Halees, 2015. An approach for detecting spam in Arabic opinion reviews. Int. Arab J. Inf. Technol, 12(1).

Aggarwal, K.K., Y. Singh, P. Chandra and M. Puri, 2005. Bayesian regularization in a neural network model to estimate lines of code using function points. J. Comput. Sci., 1: 505-509.

Al-Batah, M.S., M.S Alkhasawneh, L.T. Tay, U. Ngah and H.H. Lateh, 2015. Landslide occurrence prediction using trainable cascade forward network and multilayer perceptron. Math. Problems Eng., 2015: 1-9. DOI: 10.1155/2015/512158

Al-Batah, M.S., N.A.M. Isa, K.Z. Zamli and K.A. Azizli, 2010. Modified recursive least squares algorithm to train the Hybrid Multilayer Perceptron (HMLP) network. Applied Soft Comput., 10: 236-244.

Al-Batah, M.S., N.A.M. Isa, K.Z. Zamli, Z.M. Sani and K.A. Azizli, 2009. A novel aggregate classification technique using moment invariants and cascaded multilayer perceptron network. Int. J. Mineral Process., 92: 92-102.

Alhazmi, M. and N. Salim, 2015. Arabic opinion target extraction from tweets. ARPN J. Eng. Appl. Sci., 10(3).

Alkhasawneh, M.S., U. Ngah, L.T. Tay, M.S. Al-batah and N.A.M. Isa, 2013. Determination of important topographic factors for landslide mapping analysis using MLP network. Sci. World J., 2013: 1-12.

Alkhasawneh, M.S., U. Ngah,, L.T. Tay, M.S. Al-batah and N.A.M. Isa, 2014. Intelligent landslide system based on discriminant analysis and cascade-forward back-propagation network. Arab. J. Sci. Eng., 39: 5575-5584.

Al-Smadi, M., O. Qawasmeh, B. Talafha and M. Quwaider, 2015. Human annotated Arabic dataset of book reviews for aspect based sentiment analysis, Presented at the 3rd International Conference on Future Internet of Things and Cloud (FiCloud), pp. 726–730.

Al-Subaihin, A.S. and H.S. Al-Khalifa, 2014. A System for Sentiment Analysis of Colloquial Arabic Using Human Computation. The Scientific World Journal, 631394. http://doi.org/10.1155/2014/631394.

Aly, M.A. and A.F. Atiya, 2013. LABR: a large scale Arabic book reviews dataset, Presented at the ACL (2), pp. 494–498.

Baareh, A.K., A.F. Sheta and M.S. Al-Batah, 2012. Feature based 3D object recognition using artificial neural networks. Int. J. Comput. Applic., 44: 1-7.

Badaro, G., R. Baly, H. Hajj, N. Habash and W.A. El-Hajj, 2014. large scale Arabic sentiment lexicon for Arabic opinion mining. ANLP 2014;165:2014.

Cybenko, G., 1989. Approximation by superpositions of a sigmoidal function. Math. Control Signal Syst., 2: 303-314. DOI: 10.1007/BF02551274

Dhande, L.L. and G.K. Patnaik, 2014. Analyzing sentiment of movie review data using naïve bayes neural classifier. Int. J. Emerg. Trends Technol. Comput. Sci., 3: 313-320.

Duwairi, R.M. and I. Qarqaz, 2014. Arabic sentiment analysis using supervised classification. Proceedings of the International Conference on Future Internet of Things and Cloud, Aug. 27-29, IEEE Xplore Press, Barcelona, Spain, pp: 579-583.

Duwairi, R.M., 2015. Sentiment analysis for dialectical Arabic. Proceedings of the 6th International Conference on Information and Communication Systems, Apr. 7-9, IEEE Xplore Press, Amman, Jordan. DOI: 10.1109/IACS.2015.7103221

Elarnaoty, M., S. AbdelRahman and A. Fahmy, 2012. A machine learning approach for opinion holder extraction in Arabic language, ArXiv Prepr. ArXiv:12061011.

El-Beltagy S.R. and A. Ali, 2013. Open issues in the sentiment analysis of Arabic social media: a case study, Presented at the 9th International Conference on Innovations in information technology (iit), pp. 215–220.

El-Beltagy, S.R. and A. Ali, 2013. Open issues in the sentiment analysis of Arabic social media: A case study. Proceedings of the 9th International Conference on Innovations in Information Technology, Mar. 17-19, IEEE Xplore Press, Abu Dhabi, UAE.

El-Makky, N., K. Nagi, A. El-Ebshihy, E. Apady and O. Hafez et al., 2015. Sentiment analysis of colloquial Arabic tweets. The 3rd ASE International Conference on Social Informatics (SocialInformatics 2014) At: Harvard University, Cambridge, MA, USA.

ElSahar, H. and S.R. El-Beltagy, 2015. Building Large Arabic Multi-Domain Resources for Sentiment Analysis. In: Computational Linguistics and Intelligent Text Processing, Gelbukh, A. (Ed.), Springer, Cham, pp: 23-34.

Isa, N.A.M., M.S. Al-Batah, K.Z. Zamli, K.A. Azizli and A. Joret *et al*., 2008. Suitable features selection for the HMLP and MLP networks to identify the shape of aggregate. Construct. Build. Mater., 22: 402-410. DOI: 10.1016/j.conbuildmat.2006.08.005

Leung, K.M., 2018. Naive Bayesian classifier.

Mountassir, A., H. Benbrahim and I. Berrada, 2012. An empirical study to address the problem of unbalanced data sets in sentiment classification, Presented at the IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 3298–3303.

Mourad, A. and K. Darwish, 2013. Subjectivity and sentiment analysis of modern standard Arabic and Arabic microblogs, Presented at the Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 55–64.

Nabil, M., M. Aly and A. Atiya, 2014. LABR: A large scale arabic sentiment analysis benchmark.

Nabil, M., M. Aly, A.F. Atiya, 2015. ASTD: Arabic sentiment tweets dataset, Presented at the Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2515–2519.

Pang, B., L. Lee and S. Vaithyanathan, 2002. Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the Conference on Empirical Methods in Natural Language Processing, (NLP' 02), Association for Computational Linguistics Stroudsburg, PA, USA, pp: 79-86.

Pontiki, M., D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos and S. Manandhar, Semeval-2014 task 4: aspect based sentiment analysis, Presented at the Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pp. 27–35.

Radovanovic, M. and M. Ivanovic, 2018. Text mining: Approaches and applications. Novi. Sad. J. Math., 38: 227-234.

Refaee, E. and V. Rieser, 2014. An Arabic twitter corpus for subjectivity and sentiment analysis. Proceedings of the 9th International Conference on Language Resources and Evaluation (LRE'14), European Language Resources Association, Reykjavik, Iceland.

Rushdi-Saleh, M., M.T. Martín-Valdivia, L.A. Ureña-López and J.M. Perea-Ortega, 2011. OCA: Opinion corpus for Arabic. J. Am. Society Inform. Sci. Technol. 62: 2045-2054. DOI: 10.1002/asi.21598

Rushdi-Saleh, M., M.T. Martín-Valdivia, L.A. Ureña-López and J.M. Perea-Ortega, 2011. OCA: Opinion corpus for Arabic, Journal of the American Society for Information Science and Technology, 62(10): 2045-2054.

Schaffer, C., 1993. Selecting a classification method by crossvalidation. Machine Learn., 13: 135-143.

Shimodaira, H., 2018. Text classification using naive bayes.

Shoukry, A. and A. Rafea, 2012. Sentence-level Arabic sentiment analysis. Proceedings of the International Conference on Collaboration Technologies and Systems, IEEE Xplore Press, Denver, CO, USA. DOI: 10.1109/CTS.2012.6261103

Wahsheh H.A., M.N. Al-Kabi, and Alsmadi I.M., 2012. A link and content hybrid approach for Arabic web spam detection. Int. J. Intell. Syst. Appl., 5(1):30.

Wahsheh, H.A., M.N. Al-Kabi, I.M. Alsmadi, 2013. SPAR: a system to detect spam in Arabic opinions, Presented at the IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), pp. 1–6.