Environment Conservation Journal , 16 (SE) 283-291, 2015
ISSN 0972-3099 (Print) 2278-5124 (Online)
Abstracted and Indexed

Provide a data mining algorithm for text classification based on text content emotions using neural network

Ebrahim Haydari[1], Amir Reza Estakhrian Haghighi[2]

Received: 20.08.2015 Revised:25.09.2015 Accepted: 30.09.2015

ABSTRACT

One of the newest areas of research is in data mining and text mining is automatic discovery of knowledge from semi-structured text. an important application of data mining in the classification texts.neural networks have emerged as a powerful tool in the classification of the content of texts and a promising alternative to conventional methods of classification. in this study, by combining and improving the regulatory procedure parameters weighted binary artificial neural network model, we will provide an algorithm to classify content. This algorithm content of the texts of the Persian texts will be based on the polarity of emotion suitable performance and high accuracy .The proposed algorithm is implemented using the simulation software MATLAB and evaluated collected over 1200 comments in Farsi in the real environment. The results above show that the proposed algorithm is a neural network classification accuracy of 96% negative polarity and positive polarity sentences based on document content.

Keywords*: *data mining, text content classification, sentiment analysis, neural networks

Introduction

Knowledge discovery in databases was introduced for the first time by Fayyad at the first International Conference on Data Mining knowledge discovery in 1995 was held in Montreal, which express relationship Several analysis techniques Step pay to extract previously unknown knowledge from available data. In general, knowledge discovery in databases is the process of finding useful information and patterns of data and Using data mining algorithms to find useful information in the process of knowledge discovery in databases (Myrzavnd et al,2014).

There are various definitions for knowledge discovery or knowledge discovery in databases. Data mining is the process of identifying the patterns in the database understandable, useful, new and valid data. The goal is to find patterns and relationships hidden in the data. Among the properties that can be used to measure the quality of patterns found in the data are the ability of human understanding, validated by statistical criteria, novelty and usefulness. Knowledge discovery in databases can be thought of as a process by several processing steps can be defined. These steps should be applied in order to extract useful patterns in data sets (Myrzavnd et al, 2014). Data mining is a bridge between science, statistics, computer science, artificial intelligence, pattern recognition and machine learning data.

Data mining is a complex process in order to identify patterns and correct models, new and potentially useful patterns in large amounts of data in ways that are understandable and models for humans (Han,2006). Data mining with neural networks have been successfully applied to a variety of real-world classification tasks in business and applied sciences industry. In this paper we present an algorithm using neural networks for classification of data mining and text content based on emotions polarity sentences. In the first study described in Section 2. Then we will refer to the need for research into the theoretical foundations of Section 3 and Section 4 Previous research field of data mining and classification of content and then offer in Section 5 and 6 of the proposed algorithm and the results of the implementation of the findings of this study, data mining and classification of content on the polarity of the sentences of the document using Artificial Neural Network. Also evaluated by simulation software MATLAB are a classification of document content analysis and feature extraction of the 1,200 comments collected in real environment in Persian by the algorithm. Finally, in Section 7 we present conclusions and future work in this area. The results of the evaluation results show high accuracy in data mining algorithm text classification based on the polarity of emotions sentences text content.

2. The need for research and Problem Statement

Author's Address

Computer and Electrical Department of IAU- Sepidan Branch
E-mail: ebrahim_h@mail.com

Since the late 80s, with increasing data stored in the database more text, more traditional method s alone will not be able to provide powerful analysis and needs powerful ways to process This information because doing so is difficult and even impossible by humans. This emphasizes the need to use data mining to extract useful patterns and knowledge discovery of semi-structured documents. Data mining and knowledge discovery in databases can be useful for managers, planners, researchers, decision making and awareness of the current state of the organization. Data collection, is very large and rapidly increasing size is whom. Many of these data are collected from commercial software, websites, social networks in organizations, but the actual use of the data and detect weak and limited knowledge of them. Large amounts of data that are collected so they can be processed in different ways are needed today The extracted knowledge. The question is important:

How important not process samples? Or: Can be used to model neural network to classify and identify the polarity of Persian literature?

Given the importance of the subject and said, The aim of this study is to provide a data mining algorithm for text using neural network. That there is a very high accuracy in classifying content based on polarity and used a lot of texts in the fields of data mining, text.

3. Theoretical Foundations

Text mining and classification of content

Classification of documents to the concept of content-based text documents to one or more class predetermined; In other words, the classification of documents is a learning process in which supervised classifiers, function Documentation mapping of the domain D = {d1, d2, ..., dn} to a pre-defined set of classes C = {c1, c2, ..., cm} to form the basis of a set of training samples (classified documents) and the classification of documents based on the polarity of the texts means the content of the texts classified into three categories of positive emotions, negative and neutral. An important application of data mining in data classification. The data related to electronic media, there are some features that cute new techniques and algorithms to explore them (Karanikas et al,2012); So one of the newest areas of research in data mining, text mining. Text mining is used for automatic discovery of beneficial interest or semi-structured text. Text mining is used instead of the word "mining textual data" and also known as "discovery knowledge in the context of "sometimes. Text mining relies on finding new knowledge from text (Myrzavnd et al, 2014). It can be noted that the method of data mining and statistical classification and machine learning techniques, many of the trees The decision, the nearest neighbor method, decision tree, statistical learning (e.g Bayesian regression models and model) and so on. In the following we will refer to some of the methods offered by the researchers presented data mining.

Classification of neural networks in data mining and text content

A neural network is composed of several neuron activated when needed and computing done on it. In other words, these neurons are formed in connection with the problem solving process in relation to each other. Nodes that are in the input layer neuron that there is no any action on them In fact, in calculating the number of layers are also not covered. The nodes of the output layer neurons are responsive appears that the problem in their response. With hidden neurons between input and output neurons that are also in the hidden layer of the network.

Neural networks have a high ability, including the ability to learn, ability to generalize the systems, network robustness against error. Neural networks have emerged as a classification tool. Neural networks with suitable network structure controls the association or affiliation between input variables. The use of neural networks lies in some theoretical aspects. First, self-adaptive neural network-based method that they can not adjust their data without any data, explicit specifications of distribution and application form for the basic model. Second, the neural network model are nonlinear models that are complex to produce model real-world relationships. After an estimate of the hidden possibilities that is the basis for the classification rules and statistical analysis (Mahnaj ,2002). On the classification of documents based on key characteristics that can be extracted from the documents. It is clear that the recovery process is to identify and extract important role in enhancing the performance characteristics of documents classifiers Documentation. Whatever it is better properties extracted from documents, will improve the performance and efficiency of classification (Karanasou et al, 2015).

4. Research history

Classification methods, statistical and machine learning techniques have been proposed in recent years, including trees The decision, the nearest neighbor method, statistical learning (Such as regression models and Bayesian model) and machine neural networks, support vector and pointed. Bayat and colleagues (Bayat *et al*,2010) examined the use of support vector space methods and artificial neural networks to classify Persian texts using Hamshahri. The proposed procedures, using the weight assigned to the importance of words in documents. Vector space model requires high dimensional vector space.Using neural network to classify Persian documents in the vector space model has a better performance.

In (Dehbashyan et al,2010) MLP neural network trained to classify data Using gravitational search algorithm. Then used it to classify five reference datasets. The results showed that in most cases by GSA in two phases: training and testing neural network capability better correct classification data. The unique feature of GSA Method was developed and produced with sustainable answers relatively higher. This feature is used to classifiers, especially when they are used in critical applications such as medical or military, is very valuable and increase the efficiency of the security classification system response.

As well as in research (Mirdamadi et al, 2014) presented a method using networks Expressions neural segmentation Persian texts for search engines. The 4-phase algorithm using single words and two words of the existing possibility body and carefully carried 89.6% of Dhdml segmentation. This can be better segmented phrases to create a relative improvement in the efficiency of conventional methods. Reference (Ali Mardani et al, 2015) provides a method for combining Persian dictionary SentiWordNet monitoring algorithm and Support Vector Machine (SVM). In fact, the word is a set of algorithm parameters SVM. Among the hypotheses to achieve the best results, most appropriately dedicated to hypothesis product of the polarity in the number of repeated words .dqt algorithm is 83.78%.Based on the research conducted at (Cassinelli et al,2009) one of the simplest methods for weight feature weighting method binary in which the weight of each feature tk in document CD will be based on or lack of specificity equal to 1 or 0 in the corresponding prepares and documents. The method is not never intended frequency characteristics of the frequency characteristics of the document and only Details of the document to determine the presence or absence of weight. This method is used for classification algorithm based on machine learning including Bayesian and decision trees The format of the floating-point numbers for the value of the property is not acceptable. In (Rahate et al,2013) is recognized for sensitive words in a text sentiment analysis using support vector machine algorithm and Naive Bayes (NB). The sensitive words verbs, adjectives, and adverbs. This algorithm also detects positive and negative words in the text file and stored in two separate files. This algorithm was assessed for English literature and was obtained good results.

Also (Dhande *et al*, 2014) combined with Bayesian algorithm (NB) and a three-layer neural network that includes three input parameters to the first layer and use stimulate the sigmoid function could classify incoming documents into three categories: {+1, 0, -1}. The above is based neural network algorithm features detected by the NB. Has classified the above categories tape input documents with 80% accuracy. In research (Saraswathi et al, 2014) stated that Support Vector machine accuracy and better performance compared to other classification algorithms content. This is the border separating algorithm for clustering and clustering of input data. Using mathematical formulas set of points and separator page to find the data. SVM classification in the literature (eg, NB, EN) is higher quality than other classification algorithms. As a result of feature selection using TF-IDF and the use of support vector machine, the accuracy of the algorithm to classify content English texts into three categories: positive, negative, and undo is about 88%.

In this study, we present an algorithm to improve the composition and weighting of binary parameters using artificial neural network model layer (MLP), to classify content. The proposed algorithm for classification of Persian literature in comparison with other models have better performance and higher accuracy. First checked text input literally. Also remove and clean the useless words. The same operation is performed on the input text to send to the next step. Then, applying the proposed algorithm based on the extracted features, its neural network classification based on the polarity of the input text with high accuracy.

The method

Target neural networks, is trying to build models that simulate the human mind works. Neural network refers to a family of models, with a great atmosphere and flexible structure specified parameters and performance, provides an output pattern based on the input pattern presented to the neural network consists of a number of processing elements (neurons synthetic) that the neurons receive and process inputs and ultimately provide an output from it. In this study, we analyze and evaluate operations sentiment analysis and text mining, combined with improved parameters binary weighting method and artificial neural network model and classifying regulatory texts polarity will do with great precision. We formed three algorithms based on binary feature vector for each document vector weight sensitive vocabulary, sentence length document vector, vector convergence document sentences. The following will explain the process of making these three parameters.

Weight vector space emotional vocabulary words

In our model we consider a document D as a set including a document on {t1 ... tn} and specify the characteristics of the document with {f1 ... fm} and the binary algorithm to consider the presence or absence of selected characteristics of the document in each of the sentences. The rough consider the weight vector for sensitive words proposed by built-in dictionary that contains all the nouns, adjectives, adverbs and verbs in Persian literature is:

1) In the absence of the name of the show it with 0.and if there is a name, if the generic name of the show. With the number 1 and if the proper name with number 0

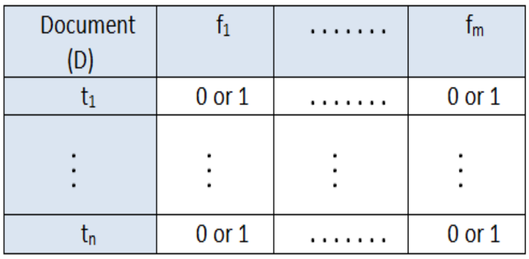
2) as represented by the number 1 and the lack of it with 0.

3) show a negative verb. Positive action with the number 0 and 1

4) indicating we display the number 1 and the lack of it with 0.

5) the existence of metaphor and simile particles represented by 1 and the lack of it with the number 0.

As a result, we form the next nm weight vector for document log. This vector is called the weight vector selected sensitive words (FS).The weight vector selection sensitive words document (FS),



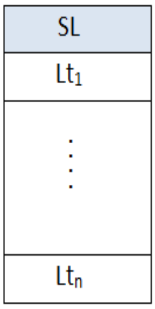
Document (D)	f_1	…….	f_m
t_1	0 or 1	…….	0 or 1
⋮	⋮	⋮	⋮
t_n	0 or 1	…….	0 or 1

Figure 1: Selected vector sensitive words (FS) document D (Source: The author, 2015)

Sentences long vector formed document:

We determined the presence or absence of emotional words based on binary algorithm.

In this phase we form the vector length of sentences in the document. Vector is the length of sentences based on the number of words in each sentence of the document. Below is Lt_1 first sentence length and the length of my Lt_n document.

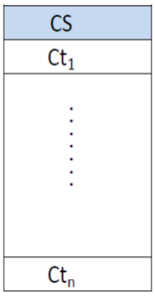


SL
Lt_1
⋮
Lt_n

Figure 2: vector length of sentences document (Source: The author, 2015)

The convergence and divergence sentences vector document:

At this point we form Vector label statements related to the input document. To form the vector consider the quickest action as identified in the following sentence, if adjectives, verbs and adjectives if they are both in the same direction, otherwise we consider the number 1 and the number 0 If there is no action as we consider the same direction.



CS
Ct_1
⋮
Ct_n

Figure 3: Vector convergence sentences document (Source: The author, 2015)

Ct_1 is the convergence and divergence of the first sentence and Ctn convergence and divergence of the n-th of the document D. We have formed this phase we selected three vector sensitive words document (FS), the vector length of sentences (SL) and vector convergence or divergence sentences (CS) document D. In fact, until this phase we have the values of the input parameters to neural networks.

In this study, the objective is to predict the classification of document content based on the polarity of sentences using the words between sentences (syntax and grammar), the convergence / divergence between (core such as type of actions) and long sentences Binary algorithms and artificial neural network model.

6. Findings

The proposed algorithm is implemented in simulation software MATLAB and evaluated with 1200 comments collected in real environment in Persian. Input data to the neural network has three steps:

1. The training set consisted of data, equivalent to 70% of the total of 70 such data.
2. Verification of information equivalent to 15% of such data 15
3. 15 information and 15% of total testing of such data.

Table 1: Summary of the implementation process (Source: The author, 2015)

Process	Percent
Examples of training	70%
Example of confirmation	15%
The collection of tests	15%
total	100%

As you can see the largest collection of education data. This set is used for training the network and obtain parameters, weight and so on. Verification of data while learning network with training data used to compare and predict. The data are provided for the measurement of interoperability testing of its network.

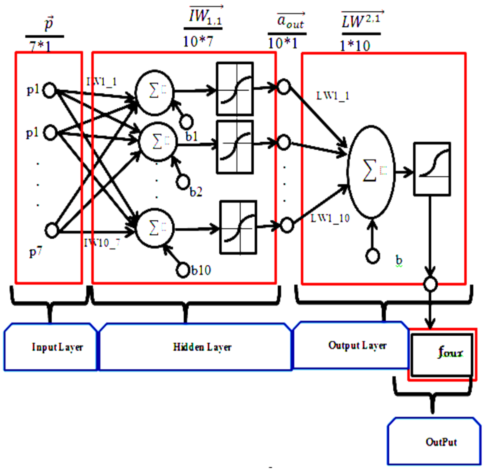

$\frac{\vec{p}}{7*1}$
$\frac{\overrightarrow{IW_{1,1}}}{10*7}$
$\frac{\overrightarrow{a_{out}}}{10*1}$
$\frac{\overrightarrow{LW^{2,1}}}{1*10}$
p1
p1
p7
IW1_1
IW10_7
b1
b2
b10
LW1_1
LW1_10
b
Input Layer
Hidden Layer
Output Layer
four
OutPut

Figure 4: The neural network is designed (Source: The author, 2015)

To determine the best network, after various design, network design, 4-layer (input layer, two hidden layers and an output layer) and 10 nodes in the first hidden layer provides the best fit. The network was presented with a layout (1-1-10-7) were used. The best function (activation function) to the layers or (subject to the weighted sum of the units in one layer and the next layer communicates values) is the hyperbolic tangent (Hyperbolic tangent). One of the main advantages of this method is that it automatically makes use of new methods of scaling up access to normalized data. The output layer activation function Log-Sigmoid. Below we show the components of the neural network architecture implementation

Table 2: Information related to network design (Source: The author, 2015)

The number of units in the input layer: UNIT 7	The number of words in	P1
	Name Type	P2
	There/ lack of quality	P3
	There/ lack living	P4
	Type of action (positive or negative)	P5
	There/ lack of a man enriched and metaphor	P6
	Convergence / Divergence (the core of the type of action)	P7
The number of the unit in the output layer: UNIT 1	Emotions polarity (positive / negative)	-
The number of the unit in the hidden layer: 2 layer and10 knots in the first layer	-	-
Subject to the active layer of hidden	Hyperbolic tangent	-

Subject to the active output	Log-Sigmoid	-
Subject Error	Uncoupling CROSS compensation too-entropy	-

The above table shows information about the neural network and is used to ensure correct allocation of cases. The multi-layer network layer in order to be connected together to form the output of the first layer and the second layer and so on that input output The last layer is the main output and up the actual response network. In other words, the network signal is fed in a direction from the input layer and leads to the output layer.

Perceptron is a non-recursive network that takes advantage of a supervised learning algorithm. Therefore, it is education category contains a set of input vectors and target vectors desire. The network includes nearly continuous input vectors of values, but the target vectors containing numbers binary ones and zeros that are generated after the training. In the following we will explain the neural network input and output functions in the hidden layer and above.

Neural network input vector features selected a binary algorithm ($\vec{p}$) and the first layer of the neural network output vector $\overrightarrow{a_{out}}$. The first layer is the best function for normalizing the output vector hyperbolic tangent function.

$$\alpha = f(W \times p + b) \qquad (1)$$

$$a_{out} = tansgn(\alpha) \qquad (2)$$

$$f(W \times p + b) = \begin{bmatrix} \mathrm{w}_{1,1} & \cdots & \mathrm{w}_{1,7} \\ \vdots & \ddots & \vdots \\ \mathrm{w}_{10,1} & \cdots & \mathrm{w}_{10,7} \end{bmatrix} \times \begin{bmatrix} \mathrm{p}_{1,1} \\ \vdots \\ \mathrm{p}_{7,1} \end{bmatrix} + \begin{bmatrix} \mathrm{b}_{1,1} \\ \vdots \\ \mathrm{b}_{10,1} \end{bmatrix}$$

(3)

Log-Sigmoid function is active function in the second layer on the output vector a_{out}:

$$\beta = f(LW^{2.1} \times a_{out} + b) \qquad (4)$$

$$\mathrm{f}(\mathrm{LW}^{2.1} \times \mathrm{a_{out}} + \mathrm{b}) = [\mathrm{LW}_{1,1} \quad \cdots \quad \mathrm{LW}_{1,10}] \times \begin{bmatrix} \mathrm{a}_{1,1} \\ \vdots \\ \mathrm{a}_{10,1} \end{bmatrix} + \mathrm{b}f_{out} =$$

$$LogSig(\beta) \qquad (5)$$

(6)

As a result, the polarity of the input is equal to:

$$Polrity(S) = \begin{cases} -1 & if \;\; 0 \le f_{out} < 0.5 \\ \\ +1 & if \;\; 0.5 \le f_{out} < 1 \end{cases}$$

(7)

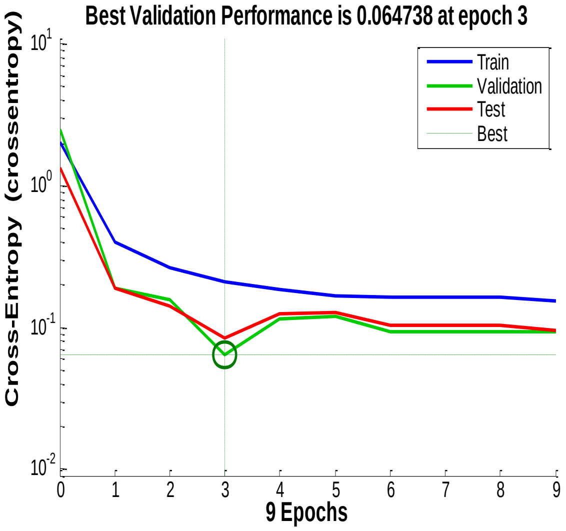


Figure 5: Chart of sample error, verification and testing (Source: The author, 2015)

when the error has not been recovered trac e amounts will continue to repeat training. The best are selected on the basis of minimizing the total error rate of repeat training data and verification

As can be seen in the above graph have fallen foul of training, verification and test iterations of network training and the network has to be a repeat optimum training (reducing errors) that AIPAC 3 the optimal value (reduce errors) is estimated to be less than 10-1.

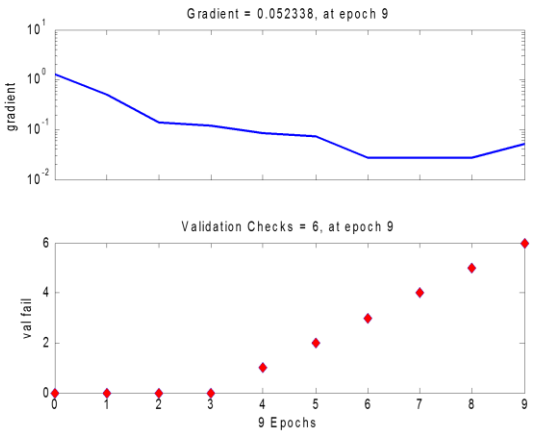


Figure 6: Network status graph (Source: The author, 2015)

The diagram above shows the status of network training, with improved error in repeated cycles, as well as the verification of compliance data with network training cycle. However, that amount is more than 0.9 coefficient R_{Square} model that is shown in the chart.

Table 3: Estimated mistake Forecast model (Source: The author, 2015)

Educational sample	Percent of forecast error	5.7%
Example confirm	Percent of forecast error	0%
Test Samples	Percent of forecast error	0%

The prediction error is expressed as the percentage derived from the classification that would follow. As can be seen in the training sample error rate is 5.7% for the other samples confirmed and there is no training prediction error.

The samples tested and approved (ie, items that have been used to build the model) sorted polarity is absolutely true feelings (without an error). There are examples of classification error as much as 5.7% as well. In total, the network is able to correctly classify 96% of the polarity of emotions.

Table 4: Summary final data (Source: The author, 2015)

Sample	Percent	CE	E%
Training	70%	0. 019394	5.7%
Confirm	15%	0. 006755	0.
Trial	15%	0. 009543	0.

Network information in the table above have been estimated. If you can see the predicted values and correct errors in sample collection and testing confirmed Shdhnzdyk to each other and generally has a very high efficiency.The neural network is more than 0.99 predictable emotions determine the polarity of the input factors. In other words, the Log-Sigmoid - which the network attempts during the process, the error rate of entropy (CE) has at least - has been successful, and in the end the predictions wrong classification table (E%) is zero. Also shown in the following diagram compliance errors each of the samples with each other.

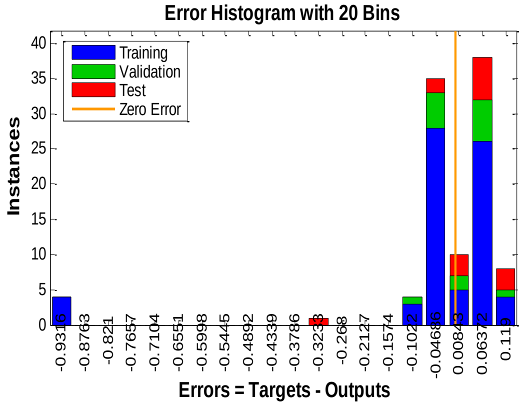


Figure 7: compliance errors each of the samples with each other (Source: The author, 2015)

As can be seen in the adaptation and reducing the amount of network errors three samples is very good.Also under the ROC curve is consistent with the provisions just in case training, verification and testing. If you can see there are only a few examples of classification error. And right there in the samples tested and approved for classified fully border and determining the polarity of emotion.

In total, the network is able to classify 96% of the correct polarity of emotions. As a result, the proposed algorithm with neural network shows classification accuracy of content on the polarity of emotion.

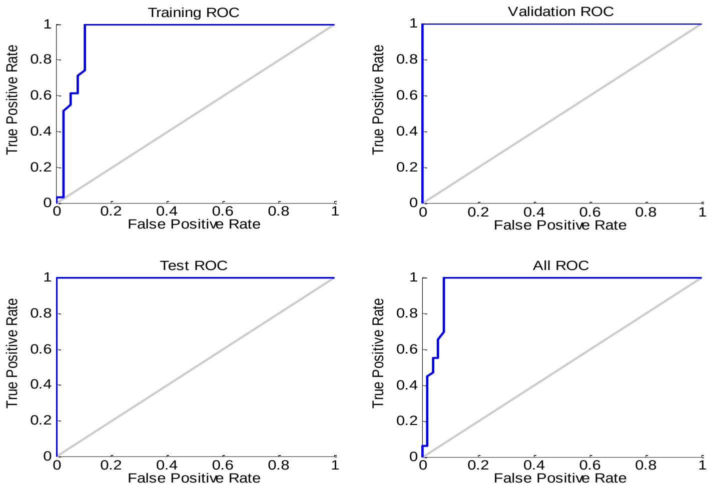


Figure 8: Chart ROC - approval provisions (Source: The author, 2015)

7. Conclusions and future work

An important method in the analysis of syntax and semantic display texts for binary algorithm. Each component features can be considered as a test and in this way we form sentences vector weight sensitive words. In addition, the proposed algorithm weight vector sensitive words using binary algorithm, we must take the words of text and vector convergence of sentences as well. Simulation and implementation of the proposed algorithm was performed on 1,200 of the comments collected in the learning environment in Persian.

The network, after various design, network design, 4-layer (input layer, two hidden layers and an output layer) and 10 nodes in the first hidden layer provides the best fit. So Arayhshd network with (1-1-10-7) and were used. The proposed algorithm in the samples tested and approved sorted feelings quite correct polarity (without an error). Educational model has 5.7% classification error as well. In total, the proposed algorithm with neural network was able to correctly classify 96% of the polarity of emotions.

Evaluation results show high accuracy in the classification algorithm and data mining, text content, including Persian texts.

The study presents a proposed algorithm using neural network model able to offer a similarly for data mining algorithms to classify text phrases polarity document. According to research conducted in this research was provided the possibility of creating a more efficient model, based on the classification of the content. It is recommended for future research topic "The combination of semantic rules sentences documents (including documents Farsi) using neural networks".

References

Ali Mardani, S., & Aghaee, A. 2015. Monitoring methods for data-mining algorithms in the Persian language using dictionaries and SVM. Tehran University, 1-18.

Bayat, M., Hosseini Khozani, S., & Gaplh, M. 2014. Automatic sorting Persian texts using neural networks, the first National Conference on Soft Computing and I T 1-6.

Cassinelli, A., & Chen, C.-W. 2009. Sentiment Categorization with Machine Learning Techniques. CS 224N Final Project, 1-12.

Dehbashyan, M., & Zahiri, S. 1389. Education MLP neural network to classify data by GSA. College of Electrical Engineering and Computer Engineering, 267-274.

Dhande, L., & Patnaik, G. 2014. Analyzing Sentiment of Movie Review Data using Naive Bayes Neural Classifier. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 313-320.

Fayyad, U., & Piatetsky-Shapiro, G. 1996. Knowledge discovery and data mining: Towards a unifying framework. In Knowledge Discovery and Data Mining. 82-88.

Han, J. 2006. Data Mining: Concepts and Techniques. University of Illinois at Urbana-Champaign-ELSEVIER.

Karanasou, M., Doulkeridis, C., & Halkidi, M. 2015. An SVM-based Approach for Sentiment Analysis of Figurative. Proceedings of the 9th International Workshop on Semantic Evaluation, 709-713.

Karanikas, H., Tjortjis, C., & Theodoulidis, B. 2012. An Approach to Text Mining using Information Extraction. Centre for Research in Information Management Department of Computation, UMIST, P.O. Box 88,Manchester, M60 1QD, UK, 1-14.

Mahnaj, M. 2002. Computational Intelligence. Center Amir Kabir University Press.

Mirdamadi, M., Zare Bideki, A., & Rezaeian, M. 1392. Expressions Persian texts segmentation using neural networks. College of Electrical Engineering and Computer Engineering, B - Computer Engineering, 76-84.

Myrzavnd, M., & Naderi. 2014. Advanced database. University Islamic Azad University research.

Rahate, R., & Emmanuel, M. 2013. Feature Selection for Sentiment Analysis by using SVM. International Journal of Computer Applications, 24-32.

Saraswathi, K., & Tamilarasi, A. 2014. Investigation of support vector machine classifier for opinion mining. Journal of Theoretical and Applied Information Technology, 291-296.