IJICTR
International Journal of Information & Communication Technology Research

Volume 15- Number 1 – 2023 (24 -34)

Identify the Subject and Content of Tweets on Twitter Using Multilayer Neural Network Method and Random Graphs

Vahid Yazdanian*
ICT Research Institute (ITRC)
Tehran, Iran
v.yazdanian@itrc.ac.ir

Mohsen Gerami
ICT Research Institute (ITRC)
Tehran, Iran
m.gerami@itrc.ac.ir

Mohammad Sadeghinia
Science and Research Branch
Islamic Azad University
Tehran, Iran
mohamad.sadeghinia.iau@gmail.com

Received: 5 May 2022 – Revised: 25 August 2022 - Accepted: 27 November 2022

Abstract—**The result of the research is a proposed model for text analysis and identifying the subject and content of texts on Twitter. In this model, two main phases are implemented for classification. In text mining problems and in text mining tasks in general, because the data used is unstructured text, there is a preprocessing phase to extract the feature from this unstructured data. Done. In the second phase of the proposed method, a multilayer neural network algorithm and random graphs are used to classify the texts. In fact, this algorithm is a method for classifying a text based on the training model. The results show a significant improvement. Comparing the proposed method with other methods, according to the results, we found that the proposed algorithm has a high percentage of improvement in accuracy and has a better performance than other methods. All the presented statistics and simulation output results of the proposed method are based on the implementation in MATLAB software.**

Keywords: Text mining, subject and content recognition, multilayer neural network, random graphs, Twitter.

Article type: Research Article

Publisher: ICT Research Institute

[DOI: 10.61186/itrc.15.1.24]

I. INTRODUCTION

In cyberspace and most social networks, it is possible to create opinions about goods, services, services, etc. It is important for a social network or website to offer its services tailored to the interests of its users. So what the user is looking for is very important. Because, it is directly effective in providing appropriate services. Therefore, reviewing and recognizing the content of texts to extract valuable information is one of the most important and challenging issues.

To analyze this voluminous and unstructured information, an extensive research field of research is presented, which is a special type of data mining. The main task of which is to analyze the texts and people's desire for a subject, existence. , Their events, phenomena, problems and characteristics are used to summarize and extract hidden and valuable

* Corresponding Author

information. In this research, we provide a suitable way to identify the subject and content of texts on Twitter.

Recognizing the topic and content of tweets is important in providing services, extracting news, and searching for related topics. Since the texts created by users on Twitter are unstructured data and cannot be directly used in building learning models, the proposed method of this research consists of several steps. In the first step, the existing data are pre-processed and the necessary features are extracted from them. In the proposed method, the dataset used for training learning models is prepared. In the next step, we use multi-layer neural network algorithm to categorize words. In this classification, we train the neural network using the random graph model and optimize the output of the network.

Using the proposed method, we increase the classification accuracy and minimize the decision error. In this research, first, materials to express more familiarity with the subject of research, then we will have a review of articles and research in this field. The proposed research method and its steps are described. In the following, the obtained results are expressed, explained and evaluated. At the end, conclusions and future suggestions are presented. All the presented statistics and simulation output results of the proposed method in this research are based on implementation in MATLAB software.

II. Problem Statement

The main idea of this research is to distinguish the content and subject of tweets from the words in them. For this purpose, we first collect the appropriate data set. This data contains words that need to be categorized correctly. Using the multilayer neural network method, we classify these words into appropriate classes. This network consists of an input layer, a hidden layer and an output layer. To classify, we give part of the data to the training department and evaluate with another part of the data. In this classification, we train the neural network and optimize the network output using a model based on random graphs. The scope of this research includes Twitter and tweet records.

III. Research Questions

Is identifying the subject and content of the texts effective in providing appropriate services to users?

Is it possible to make a good decision about the content of the texts using the semantic load of the words?

Is the multi-layer neural network method a good way to identify the subject and content of texts and their classification?

Can using a combination of multilayer neural network method and random graph model to identify the subject and content of texts increase the accuracy of classification?

Can the network output be optimized using a random graph model?

IV. Neural Network

It is a calculation method that builds several processing units based on the interconnectedness. The network consists of an arbitrary number of cells or nodes or units or neurons that relate the input set to the output [1]. Artificial neural networks are new systems and computational methods for machine learning, knowledge display, and finally the application of knowledge obtained to maximize the output responses of complex systems [2].

There is no exact agreement on the definition of the neural network among researchers; But most agree that the neural network consists of a network of simple processing elements (neurons), which can exhibit a complex set of general behaviors of the relationship between processing elements and element parameters.

The main and inspiring source for this technique comes from the experiment of the central nervous system and neurons (axons, multiple branches of nerve cells and junctions of two nerves), which is one of the most significant elements of information processing in the nervous system. Give. In a neural network model, simple nodes (broadly neurons, neurons, "PEs" (processing elements), or units are connected to form a network of nodes, hence the term It is called "neural networks" [3].

Using programming knowledge, a data structure can be designed that acts like a neuron. Then, by creating a network of these interconnected artificial neurons, he taught them how to create a training algorithm for the network and apply this algorithm to the network [3].

V. Random Graph Model

It is a model based on random graphs that provides an algorithm for generating free-scale networks using the growth characteristics and preferential connection of real networks. This algorithm starts with a complete graph with m_0 vertices and in each step a new vertex is added to the graph which is connected to the vertex of the previous vertices. The probability of an edge between the new vertex and the old vertex i is proportional to $\frac{d_i}{\sum d_i}$, ie the degree of higher nodes increases more than other nodes. As a result, the distribution of degrees remains constant over time and of the law-power type [2]. In this model, we first start with a network of N vertices whose vertices are numbered. First we connect each node i to the vertices $i \pm ۱,۲, \ldots, \frac{k}{۲}$. Then, with a probability of p, we select an edge and change one of the two nodes connected to it [2].

VI. Research Background

Twitter is a growing social network. Many companies and institutions use Twitter to brainstorm their customers. In 2017, Jianqiang et al. Analyzed Twitter comments [4].

The number of tweets and the way messages are distributed are in 3 categories. The results show 15 popular hashtags among Twitter users. In this table, in addition to hashtags, their frequency of repetition as

[DOI: 10.61186/itrc.15.1.24]

well as synonymous hashtags, if any, are given. The F-measure, which is an evaluation criterion in data mining algorithms and is also called the F-criterion, is an important criterion for measuring the number of true samples in a class that is specific to Shows the self-categorized to the whole sample, for this experiment it shows that feature extraction in n-gram combination mode is actually a method that breaks down a sentence into interconnected words. Each of these connections is considered a gram, and Lexicon and Micro-Blogging are the best answers. It has also been observed that in HASH mode the results are better than in HASH + EMOT mode. Also, the degree of classification accuracy has provided similar results, with the difference that HASH + EMOT has provided better results than HASH with a slight difference [4].

In another study conducted in 2019, Prach et al. [5] proposed a new method. The experiment was conducted in two different languages, Arabic and English. In English, POS and n-gram properties were used, and in Arabic, word root properties were used for extraction. The root of words requires a rich dictionary for recognition. The dimension reduction step and feature selection are also done automatically.

[6] Refers to the text analysis work done on the virtual store. The results were evaluated with three criteria of reading accuracy and F1-measure. This paper also examines the accuracy of analysis by machine learning for customer satisfaction.

In Mohbey's article [7] is used to analyze the extraction of law. Evaluation of these texts is measured by three criteria of accuracy, recall and F1-measure. As can be seen, for positive and negative texts, the accuracy of this algorithm is higher than other criteria.

The results of this experiment are also compared with another experiment used in the same way on another dataset. This dataset is called 20newsgroup, which includes 20 separate groups, 750 teaching texts, and 250 experimental texts in each group. In this experiment, the lowest value for F1-measure was 2.0 and the highest value was 8.0, while in BBS data the maximum value was around 7.0. The reason for this decrease is explained in the report that there is no standard in the BBS text data and the determination of training and test data has not been done by linguists and as a result there may be errors in them [7].

In [8] he has used another method with supervision. According to the author, this method gives better results than the L-HMM method. The L-HMM method is an unsupervised method that uses a large number of nodes in the hidden layer. These nodes are trained at each stage and the best decision makers are selected as the main node.

Each customer goes through three steps to buy their products from virtual stores [8]:

1- A quick look at all the products to get information such as price range, variety, etc.

2- More detailed review of each product, which includes reviewing the opinions of other consumers.

3- Product selection

According to these steps, it can be said that step 2 plays a key role in user selection. Two important factors play a key role in reviewing any product: product features and user feedback. The results show that the four factors of the general image of the society of the product, consumption, statistical specifications and product descriptions play a role in customers' purchases.

The criteria used, as in the previous sections, are call, accuracy, and F-measure for two sets of data that have relatively similar results. These criteria are implemented on two cameras based on feature comments, physical component comments, functional comments, and general product reviews.

In methods that have a semantic approach, the use of dictionary resources is a must. Because the basis of these methods will be the emphasis on the semantic and grammatical features of texts to recognize it. At present, these methods are not widely used because they are always faced with a shortage of resources. In the comparisons made between machine and semantic learning methods, it is generally understood that machine learning methods require more time to train the desired model, but on the other hand, better performance accuracy than the method. Have semantic meanings. Semantic methods have less detection accuracy, on the other hand, do not require time to build the model and will be more useful in real-time applications.

It has been used for analysis based on the semantic approach in [9]. Modes can be as follows: adjective + noun, adverb + adjective, adjective + adjective, noun + adjective, adverb + verb (note that this combination is for English grammar).

It can be seen that out of 71 positive comments, 67 positive comments were recorded correctly (77.91% positive call percentage) and out of 29 negative comments, 10 negative comments were correctly reported (71.43% negative call percentage). . In total, the accuracy of this classification is reported to be 77%.

Of course, in the next article we will point out that this difference may be related to the structural differences of education and for each corresponding structure of education it includes a better result.

To improve the results, the generated texts are given sentence by sentence to categories that work based on semantic properties. One method is to compare each node on the tree with its subdivisions to see at what point the similarities and differences are best. The accuracy of this method sometimes exceeds the accuracy of the n-gram method. In this article, it was decided not to repeat these steps until a high number and to set a threshold for it. Another challenge is what criteria to use to evaluate this method.

To improve this method, some suggestions on how to streamline data are given below. Among the existing algorithms for streamlining the NB algorithm, the best answer was included, but this algorithm was also tested along with other methods. The best result was NB obtained from combination with Laplace and improved the result up to 87% [10].

[DOI: 10.61186/itrc.15.1.24]

BERT is a new pre-trained language representation model published by Google to represent words. BERT is free and NLP practitioners can use it to create models. BERT has high performance in various languages understanding benchmarks and it captures structural information about language [13].

BERT is a new language representation model, which achieved the state-of-the-art performance on most of the NLP tasks it has been applied upon. Moreover, it outperforms number of traditional NLP techniques [14]. Most of text classification studies based on BERT showed strong results in several languages. Authors in [15] developed a classification solution based on BERT-model, in order to detect self reports of prescription medication abuse from Twitter. They reported that their proposed model performed better than the best traditional model. In [16], authors performed an application of sentiment analysis applied on Italian tweets according to their polarities.

BERT pre-trains using both masked sentence modelling (MLM) and next sentence prediction (NSP) techniques [17]. MLM will choose random words and then conceal them. The model is then expected to predict the hidden words; this helps to avoid the issue in bidirectional models where the words can see themselves [17]. NSP provides the model with paired sentences to see if the model can predict if the second sentence comes after the first or not; this allows the model to learn how sentences interconnect [18].

VII. STEPS OF THE PROPOSED METHOD

A categorization model consists of categories that combine the opinions and decisions of each member of the group to categorize new examples. This combination of the views of individual categories is done in the classification issues by voting.

In this research, a multilayer neural network algorithm has been used for this purpose. In fact, the issue we are investigating is a data mining issue of the category that aims to identify the content of tweets to provide services tailored to the user's request. Figure 1 shows a diagram of the proposed method.

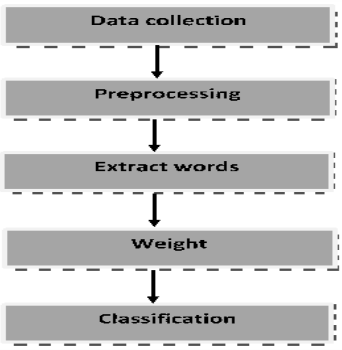

Figure 1. Steps of the proposed method.

First, the data is collected from Twitter, then it is pre-processed and the best data set is selected to teach the learning model. The sentences are broken down into their constituent words, and according to the word scoring table, each word is given the desired weight. This weight vector is given to the neural network algorithm and the network weight is optimized in each iteration using the Watts-Strogatz model and a training model is constructed.

According to Figure 1, preprocessing must be performed to construct a new data set that can be used to teach the learning model. After creating a new data set, the main step, namely teaching the proposed learning model in this research, which is the neural network, is performed and the Watts-Strogatz model is used to weigh the network.

VIII. DATA PRE-PROCESSING

Data preprocessing is actually one of the stages of the data mining process. The cornerstone of a good data mining operation is the use and access to good and appropriate primary data, which is referred to as data preparation or pre-processing. To pre-process the available data, we have used the conventional pre-processing method that has been used in many data mining works related to social networks, which is also used in the article of Shaima et al. All the pre-processing part is done in MATLAB software.

Data preparation spends about 21-31% of the time required for data mining and 17-31% of the success of data mining projects is related to it. Failure to prepare data or its poor preparation causes the complete failure of the project. Data preprocessing is necessary to improve the quality of real data for data mining. A quality data is read that is correct, complete, consistent, up-to-date, acceptable, valuable, interpretable and accessible. Data pre-processing is an important step in the direction of successful data mining. . The actions that are performed in data preparation are [10]:

- Clearing data
- Data integration
- Data conversion
- Data reduction

Based on the type of application on which the data mining operation should be performed, different techniques are used for each of these operations. The figure below shows a view of the main pre-processing components.

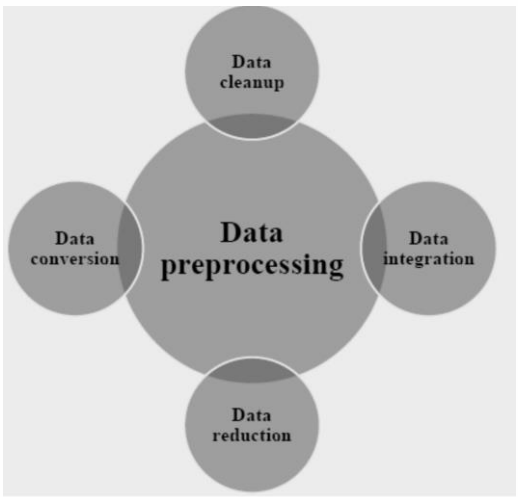

Figure 2. The main components of data preprocessing.

[DOI: 10.61186/itrc.15.1.24]

The data pre-processing stage includes the following steps [10].

Removing additional symptoms:

The first phase includes data pre-processing. First, the existing data set, which includes texts, is loaded into the memory. Then, the textual data is cleaned. Numbers, internet addresses, signs and meaningless characters are removed from the text. Then, the texts that are in the form of sentences are converted into words and each text becomes an array of words.

Marking and removing less important words:

In the following, frequent and unimportant words are removed from the text that are broken down into their component words. Then the features are extracted and assigned to these features using notation. which is actually the formation of ngrams of each sentence as a feature and their value based on the bag of words method. In the following, it values its features for each text using markup.

Mapping to feature space:

Finally, the output from the previous step will be a matrix in which each row represents a text and the columns represent the features. This matrix maps the data set to a new feature space in which each row contains text and each column contains features related to that text. This matrix is used to teach learning models in the second phase.

Feature extraction and selection in order to extract and weight words:

There is a lot of interest in feature extraction, feature building and feature selection among experts in the fields of statistics, pattern recognition, data mining and machine learning; One of the goals of this work is to predict and detect errors.

Even with the current advanced computer technology, discovering information from data is still a very difficult task due to the characteristics of computer data. Feature extraction, construction, and selection are a set of techniques that simplify and transform data, thus making diagnosis tasks easier.

Feature selection can be used to simplify the program language. In some other cases, the language used may not be sufficient to describe the problem for some learning algorithms. In these cases, the feature building trick can be used to enrich the language used. Of course, sometimes these built-in features are not all useful. Feature selection can later remove and remove these unusable features. It is also very common to see feature extraction and feature selection merged.

As computer and database technologies develop at a significant speed, humans rely more and more on computers to gather data, process data, and use data. Machine learning, information discovery, and data mining are some of the smart tools that help humans achieve these goals.

Researchers and practitioners have found that to use these tools effectively, an important part of the work is preprocessing, where data is processed and prepared before being fed to any learning, discovery, or simulation algorithms. be made

In many applications of discovery, a key factor is finding a subset of the population under study that behaves enough like that population to be worth focusing our analysis on. Although most learning methods try to select and extract or build features, both theoretical analysis methods and experimental studies state that many algorithms in a wide range of unrelated features or additionally, they work poorly. All evidence points to the need for additional methods to overcome these problems.

Data transformation and selection are among the techniques that are widely used in the pre-processing process. Data transformation is a process during which a new set of features is created. The types of data transformation are data creation and data extraction. Sometimes, both of these cases are known as feature discovery.

Feature creation is a process that explores information about the relationships between features and strengthens the space of features by inferring or creating additional features.

Feature extraction is a process that extracts a new set of main features by using some basic and functional routing. Using feature extraction, we can determine the salient and defining features of the data.

Subset selection differs from feature change in that no new features are produced in the subset selection, and only a subset of the main features are selected and the space of features is reduced.

Similar to the feature modification process, feature construction typically expands the feature space, while feature extraction typically reduces the feature space. Changing the features and selecting the subset are not completely separate and independent from each other. They can be considered as two different sides of the problem representation.

In the proposed method, each sentence is broken into its component words and each word is assigned a corresponding weight. The result of this work is the formation of word weight vectors (feature vectors) which are used as input to the classification learning model.

IX. CLASSIFICATION OF TWEETS

After performing the pre-processing phase and extracting feature vectors, it is time for the main stage of the work that is, categorizing tweets. In this research, multilayer neural network algorithm has been used for this purpose. In this algorithm, the training model is built first, and after the training is finished, the training samples are used to classify a new sample. Therefore, using the data created in the first step, which are the feature vectors, we train and evaluate the neural network model in the following way.

One of the efficient tools used in problems related to classification or estimation (prediction, regression) of the target variable is neural network. This method is a calculation method based on the interconnected connection of several processing units.

The network consists of an arbitrary number of cells or nodes or units or neurons that connect the input set to the output. In this section, a type of neural network

[DOI: 10.61186/itrc.15.1.24]

called perceptron is used. A perceptron takes a vector of inputs with real values and calculates a linear combination of these inputs. If the result is greater than a threshold value, the output of the perceptron will be equal to 1, otherwise it will be equal to -1. To learn network weights, Back Propagation method is used.

In this method, by using gradient descent, it is tried to minimize the square of error between the outputs of the network and the objective function. In fact, in order to solve the problem in general, the input modules are converted into a set of features and the network is trained based on these features. At the end, the network selects the format that best matches the input.

By increasing the size of the network, the accuracy on the training samples increases. But the accuracy of detection has decreased on other data. In fact, when the size of the network increases, the complexity of the network decreases the accuracy during testing, while the accuracy continues to increase in the training samples. When there is an error or noise in the training data, the network grows and becomes more complicated due to the presence of the error in order to adapt to all the training data. On the training data, the accuracy is high, but during the test, its accuracy decreases.

Watts-Ostrogatz model is used to determine the weight of the network. so that first each word vector graph starts with m_0 vertices (N numbered vertices) and at each step a new vertex is added to the graph (every node i is added to vertices i±1,2,..., k/2) which is connected to the vertex from the previous vertices in such a way that we select an edge with probability p and change one of the two nodes connected to it, and the degree of higher nodes is more than the node others increase. This causes important words with more weight to play a greater role in classification.

X. EXTRACT TRAINING DATA

In this step, the mapping of the obtained features is used. Part of the data is randomly selected as training data and other data is used for evaluation.

XI. NEURAL NETWORK TRAINING AND WEIGHTING BASED ON WATTS-STROGATZ MODEL

At this stage, the neural network learning model is constructed, weighted by the Watts-Strogatz model, and used to classify texts. In this way, the properties quantified in the previous step, as input input vector, are given to the training and neural network model creation, and after creating the learning model, each time using the Watts-Strogatz model, the network weight is optimized. To be.

To obtain random graphs whose nodes and links are random, we used the Watts-Strogatz model, whose inputs are the row and column matrices of the mapping, which are the number of input layer neurons and the number of lattice hidden layers, respectively. they will be.

The output of this function will be a matrix W with dimensions row * col, which is actually the final matrix that we will apply to the weights of our trained network. It will also connect the input layer nodes to the hidden layer neurons, which will be used to draw the links.

There are at least d + 1 nodes in the graph with an average of d and more links, and older nodes have more links, which is derived from the preference rule in the Watts-Strogatz model.

The main reason for combining this method with artificial neural networks in this study is the interesting theory about the Watts-Strogatz model. Existing theories show that when establishing communication between neurons in the brain, in addition to storing memory information, etc., a certain minimum number of connections between neurons are needed, in addition to the age, age and experience of these neurons when communicating. Previous neurons are also important. This is also true of human social networks that follow the theory of complex systems.

The proposed method uses the Watts-Strogatz model in neural network weighting. This model is a model in which the majority of nodes are not neighbors but can communicate with each other through a small number of intermediaries. This feature is known as "six degrees of separation". This model has a low path length and high clustering coefficient that is seen in many real world networks such as social networks, Internet and genetic networks. We have used this feature in the neural network and combined this model with the neural network. In fact, we have created a neural network based on the Watts-Strogatz model. The Watts-Strogatz model, which is one of the main models for creating small world networks, has provided the ability to make the best available category by considering the least feature in the category. In this way, each feature is weighed according to its importance in decision-making and is connected to other nodes.

In this research, a six-layer neural network has been used, which will have a diagram in the form of the following figure.

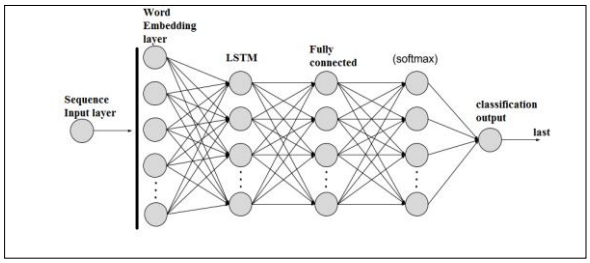

Figure 3. An overview of the designed neural network.

After all the sentences in the dataset have been broken down into their constituent words, 70% of these words, which are actually attribute vectors, are taught to the department to build a model based on it, which contains 22 million words. . The remaining 30% is also used to test the model. The initial layer of data entry is encoded.

We then deliver the encoded information to a pre-trained network that has trained 22 million words and activate 1000 layers along the total vocabulary of the word bag in which the weights will be a matrix of 22 x 1000, the output of this step. Which is a 1000 vector as input to the LSTM network. The third layer will be the LSTM network. With an input dimension of 100 that is transmitted to 2200 neurons. So we have the input

[DOI: 10.61186/itrc.15.1.24]

weight matrix with dimensions of 2200 * 100, then we connect these 2200 neurons to 1100 neurons in the hidden layer and the weight matrix of 2200 * 1100 middle layer of LSTM is obtained.

The fourth layer of a network will be fully connected to the number of classification classes. Whose weight matrix will later accept 2 *1100.

Then we pass the output of this step through a softmax network and report the classification results with the last layer.

XII. Classification based on training model

The training model was created using the previous steps. In this step, the classification is done based on the created model. So that each new input sample is processed and the steps from the beginning of the season until now are performed on it. Then the degree of similarity with the classes created in the model is examined. Depending on the degree of similarity, each class in which the input sample has the highest degree of similarity is identified as the class of that sample.

XIII. Data used in the proposed method

The data used in this study was extracted from the Twitter repository. This dataset contains 12800 records. In this dataset, each record is written by a user and has a specific tag that indicates the subject of the text. The available data are presented in the form of xls files, for all of which we have created xml files, which after initial processing and loading it into a format usable for processing, enters the proposed algorithm as input to the first phase. To be. Each sentence is placed in a sentence tag. By processing xml files with the help of MATLAB libraries, the texts are sent as the primary input in the proposed method.

As previously explained, the training and assessment dataset is available offline. The dataset used included text on Twitter.

XIV. Implementation steps

Phase One: Pre-processing the initial data

The first phase involves data preprocessing. The existing data set, which includes texts, is first loaded into memory. Then, the text data is cleaned. Numbers, URLs, and signs and symbols and meaningless characters are removed from the text. Then, the hockey text, which is in the form of a sentence, becomes word for word, and each text becomes an array of words.

Repetitive and insignificant words are then removed from the text that have been broken down into their constituent words. Then the properties are extracted and these properties are quantified. In fact, the formation of ngram of each sentence as a feature and their quantification is based on the word bag method. It then values its features for each text. Finally, the output of this operation will be a matrix in which each row represents a text and the columns represent the properties. This matrix is used to teach learning models in the second phase.

Phase Two: Classification

In this phase, the multilayer neural network learning model is constructed, weighted by the Watts-Strogatz model, and used for classification. In this way, the properties set in the previous step, as input feature vectors, are given to the training section and the creation of the neural network model, and after creating the learning model, the weight of the network is optimized each time using the Watts-Strogatz model.

XV. Compare the efficiency of the proposed method with other methods

For this purpose, we compare the proposed method with a hybrid algorithm and a group algorithm, which are the best methods proposed for text mining so far.

First, to compare the proposed method with deep methods, we use the deep neural network classification algorithm presented in the article [11] and to compare the proposed method with group learning methods, we use the method presented in [12]. Deep algorithms and mass algorithms are popular and widely used data mining methods and have been widely used in studies. The results of this evaluation are shown in Tables 1 and 2.

XVI. Evaluation criteriar

Our evaluation criteria are the degree of accuracy and retrieval in the classification of classes in the data. We have compared the evaluation criteria in our proposed algorithms with other methods and we have shown the results of these comparisons in the following tables.

In Table 1, we have examined the proposed algorithm in terms of the evaluation criteria mentioned. Evaluation criteria are a key factor in measuring the performance of classification. To calculate the criteria, we use formulas 1, 2 and 3.

$$\text{Accuracy}= \frac{TP+TN}{TN+TP+FN+FP} * 100\% \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} * 100\% \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} * 100\% \quad (3)$$

TP: Positive examples that are correctly classified.

TN: Negative examples that are classified correctly.

FP: Positive examples that are incorrectly classified.

FN: Negative examples that are incorrectly classified.

In the present study, the topics are in the categories of news, sales, sports and economics, etc., and examples related to news and sales are considered as positive examples, and other samples are considered as negative samples.

TABLE I. How to identify evaluation criteria

	Description	Sample	Class detected by the algorithm	The main class
TP	Number of negative samples that are correctly classified	The magnitude of the recent earthquake is estimated 6.2.	News	News
TN	Number of positive samples that are	The Iranian Handball team qualified for the	Sports	Sports

[DOI: 10.61186/itrc.15.1.24]

	incorrectly classified	Asian Championships.		
FP	Number of negative samples that are incorrectly classified	Housing transactions are declining.	News	Buy and Sell
FN	Number of negative samples that are correctly classified	Currency prices are rising in the open Market	Buy and Sell	Economical

TABLE II. THE DEGREE OF ACCURACY AND RETRIEVAL OF THE PROPOSED ALGORITHM

Criteria for evaluating the proposed method	
Accuracy	95%
Precision	99%
Recall	86.9%
Run time in seconds	12.35

TABLE III. COMPARISON OF THE PROPOSED METHOD WITH OTHER METHODS

Method	Accuracy	Recall	Precision	Run time in seconds
Suggested Method	99%	92.2%	95%	12.35
Collective method	89%	91%	87%	16.42
Deep neural network method	78.3%	80%	87.8%	24.56

According to the results of Tables 2 and 3, it can be seen that the performance of the proposed algorithm is better and has a favorable situation in the evaluation criteria.

XVII. ANSWERS TO RESEARCH QUESTIONS

Is identifying the subject and content of the texts effective in providing appropriate services to users?

Yes. According to the efficiency of the proposed method and the results obtained from the simulation of the proposed method and its use according to Tables 2 and 3, users with 95% accuracy are able to identify their favorite topics and with less time can Get the services they are looking for, and based on their interest, topics related to their interest are suggested and provided to them.

Is it possible to make a good decision about the content of the texts using the semantic load of the words?

The results obtained by MATLAB software according to Tables 2 and 3, show that the use of semantic load of words has been a suitable method and has created an acceptable level of accuracy equal to 99% and 95%.

Is the multi-layer neural network method a good way to identify the subject and content of texts and their classification?

The results obtained from the implementation of the proposed method in MATLAB software and its testing on detection and classification on 30% of the data set according to Tables 2 and 3, indicate that the multi-layer neural network method of classification accuracy Has increased to 99% and also increased the classification accuracy to 95%. In addition, it was able to reduce computation time. Therefore, the multilayer neural network method is a suitable method for identifying the subject and content of texts and their classification.

Can using a combination of multilayer neural network method and random graph model to identify the subject and content of texts increase the accuracy of classification?

The results of implementation in MATLAB software according to Table 3, has shown that the use of a combination of multilayer neural network and random graphs to identify the subject and content of texts has been able to classify accuracy compared to the mass method of 8% and 7% improvement over deep method. It has also been able to increase accuracy to 95%.

Can the network output be optimized using a random graph model?

The results of implementation in MATLAB software have shown that the use of random graphs has optimized the network output and reduced the number of intermediate nodes. As a result, the search space is reduced and the time to reach the ideal answer is reduced according to Table 2. Also, with increasing weight, the effective properties of the layers are less complex and the classification accuracy is increased.

XVIII. CONCLUSION

The result of the research was a model for text mining on Twitter, the function of which seems to be appropriate in categorizing texts. We designed and performed experiments to evaluate the proposed method. We observed that the results show a significant improvement. Comparing the proposed method with other methods, according to the obtained results, we found that the proposed algorithm has a high percentage of improvement in accuracy and readability and has a better performance than other methods.

In general, machine learning methods will work much better than dictionary-based methods. Finally, it can be concluded that using the proposed method to categorize texts in virtual stores, sales sites, or any organization or trustee that offers a product or service online, is quite promising and with further studies Better results can be achieved.

According to the efficiency of the proposed method and the results obtained from the simulation of the proposed method and its use according to Tables 2 and 3, users are able to identify their favorite topics with an accuracy of more than 95%. The results also show that the use of semantic load of words has been a good method and has created an acceptable degree of accuracy equal to 99% and 95%. These results indicate that the multilayer neural network method has increased the classification accuracy up to 99% and also the classification accuracy up to 95%. In addition, it was able to reduce computation time. Therefore, the multilayer neural network method is a suitable method for identifying the subject and content of texts and their

[DOI: 10.61186/itrc.15.1.24]

classification. Also, the use of a combination of multilayer neural network and random graphs to identify the subject and content of texts has been able to improve the classification accuracy by 8% compared to the mass method and 7% compared to the deep method. It has also been able to increase accuracy to 95% and reduce the search space and reduce the time to reach the ideal answer according to Table 2. Also, with increasing weight, the effective features of the layers are less complex and the classification accuracy is increased.

XIX. Suggestions and Future Work

The present study is specifically focused on the classification of texts. Texts are raw information without structure that needs to be processed. To evaluate the proposed method, this model can be used in real-time text processing, which includes more numbers and resources. This proposal is presented in line with the research because the number of sources under review is high.

In future research, it is possible to expand the topics and classify them into more categories with different topics in order to study more texts and extract more accurate information from them. This suggestion can be used in other research because the data of this research include the same number of topics that have been addressed in this research.

It is also possible to use other methods instead of using the word bag method and increase the efficiency of the algorithm with the help of these weighting methods. This suggestion can be used in other research.

In this context, the dimensions of the feature space can also be reduced by using the appropriate feature selection methods. Provided, of course, that it does not adversely affect the performance of the learning algorithm. This suggestion can also be used in other research.

References

[1] Altman, M,. (1994), "Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience)", Journal of Banking and Finance, Vol: 18, PP: 505-529.

[2] Cybinski, P,. (2010), "Discription, Explanation, Prediction, the Evolution of Bankruptcy Studies", Faculty of International Business and Politics, GriffinUniversity, Brisbane, Vol:27, No:4, PP:29-44.

[3] Anandarajan, M, and et al,.(Jun 2011), "Bankruptcy Prediction of Financially Stressed Firms: An Examination of the Predictive Accuracy of Artificial Neural Networks", International Journal of Intelligent Systems in Accounting, Finance and Managment, Vol: 10, No: 2, , PP: 69-81.

[4] Jianqiang, ZH., Xiaolin, G,. Comparison research on text pre-processing methods on Twitter sentiment analysis, IEEE Access 5, 2870-2879, 2017.

[5] Proksch, s., et al,. Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches, Legislative Studies Quarterly 44 (1), 97-131, 2019.

[6] Luo, X., et al,. User behavior prediction in social networks using weighted extreme learning machine with distribution optimization, Future Generation Computer Systems 93, 1023-1035, 2019.

[7] Mohbey, K., Multi-class approach for user behavior prediction using deep learning framework on twitter election dataset, Journal of data, Information and management 2 (1), 1-14, 2020.

[8] Bao, B., Chen, L., Cui, P., User behavior and user experience analysis for social network services, Wireless Networks, 1-7, 2020.

[9] Huang, L., et al., User behavior analysis and video popularity prediction on a large-scale vod system, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 14 (3s), 1-24, 2018.

[10] Shaimaa, M., et al., Classification and prediction of opinion mining in social networks data, International Journal of Computers and Information, 2020.

[11] Sheng, X, X. Wu, Y. Luo, A Novel Text Mining Algorithm based on Deep Neural Network, IEEE, 2018.

[12] Chen, Z, B.Liu, Mining topics in documents : standing on the shoulders of big data, in: Proceedings of the 20thACMSIGKDDInternational Conference on Knowledge Discovery and Data Mining,ACM ,pp.1116–1125, 2020.

[13] Jawahar, G., Sagot, B., & Seddah, D. (2019, July). What does BERT learn about the structure of language?

[14] [14] S. Gonzalez-Carvajal and E. C. Garrido-Merch ` an, "Comparing BERT ` against traditional machine learning text classification," ArXiv preprint, 2021, arXiv:2005.13012

[15] M.A. Al-Garadi, YC. Yang, H. Cai, Y. Ruan, K. O'Connor, G.-H. Graciela et al., "Text classification models for the automatic detection of nonmedical prescription medication use from social media," BMC Med Inform Decis Mak, 2021, pp. 21-27, doi: 10.1186/s12911-021-01394-0

[16] M., Pota, M. Ventura,R. Catelli and M. Esposito, "An effective BERTbased pipeline for Twitter sentiment analysis: A case study in Italian," Sensors, 2021, vol. 21, pp. 1-133, doi: 10.3390/s21010133

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805, 2018.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems 30, pages 5998–6008. Curran Associates, Inc., 2017.

Vahid Yazdanian received his Ph.D. degree in University of Technology (Tehran Polytechnic) in the field of control of two-dimensional systems. He is an Assistant Professor at ICT Research Institute (ITRC) in Tehran. His research interests include Artificial intelligence, Block chain and Cryptocurrency, Cyber Security, Digital transformation, Information and Communication Technology and ICT Policy.

Mohsen Gerami received his Ph.D. degree in Engineering of Information and communication Technology from Seoul National University. He is an Assistant Professor at ICT Research Institute (ITRC) in Tehran. His research interests include Security, Block chain and Cryptocurrency, Cyber Security, Digital transformation, Information and Communication Technology and ICT Policy.

[DOI: 10.61186/itrc.15.1.24]

Mohammad Sadeghinia received his M.Sc. Degree in Information Technology. He is a Computer Expert. His research interests include Business Support System, Security, Cyber Security, Digital Marketing, Information and Communication Technology.

Appendix

Comparison with other works

Author and year	Jianqiang et al. 2017	Proksch, s., et al. 2019	Luo, et al. 2019	Mohbey 2020	Bao et al. 2020	Huang et al. 2018	Shaimaa et al. 2020
Title	Sentiment analysis on Twitter and comparative research on texts using pre-processing methods	A new method for multilingual sentiment analysis	Analysis of the behavior of social network users using machine learning	A multi-class approach to predict customer behavior using deep learning on Twitter	Analysis of the experiences and behavior of users of social networking services	Analysis of user behavior on video products in large-scale systems	Classification and prediction of opinions in social network data
Method	Feature extraction in combination mode of n-gram, Lexicon and Micro-Blogging	The feature of word roots and the dictionary method are used for extraction. Dimension reduction and feature selection is done automatically	Using conditional probability distribution on undirected graph model	Law extraction is used for analysis. By using this method, users can be effective in creating a meaningful space.	Used one method with another supervision. A more detailed review of each product, which includes the reviews of other consumers, is considered at each stage.	Analysis based on the semantic approach, as well as the use of semantic and thematic modeling in the analysis	Suggestions have been given regarding smoothing the data. Among the existing algorithms for smoothing the NB algorithm, the best answer is included.
Results	The amount of F-measure for this test in HASH mode has been better than HASH+EMOT mode. Also, the level of classification accuracy has provided similar results, with the difference that HASH+EMOT has provided better results than HASH with a slight difference.	By using random graphs, the effective features have been identified and as a result, the output of the network has been improved.	The accuracy of machine learning methods is high. It has higher accuracy than conventional methods such as average, weighted sum.	The detection accuracy of the method is high. The determination of the teaching and testing data was not done by linguistic experts, and as a result, there may be errors in them	The results show that the four factors of the society's general image of the product, the amount of consumption, the statistical characteristics and the description of the product play a role in the purchase of customers.	It has increased the accuracy of diagnosis. In supervised methods, the accuracy percentage has always been higher than the semantic approach. Linguistic approach methods have better results than machine learning methods	The best result was obtained by NB, which was obtained by combining with Laplace and improved the result by 87%.
Results of the proposed method in response to the ::							

[DOI: 10.61186/itrc.15.1.24]

first research question	The use of the semantic load of words has led to the correct recognition of the content.
second research question	By correctly identifying the user's interests, it provides more suitable services to the users.
third research question	The results show that the multi-layer neural network method is a suitable method for recognizing the topic and content of texts and classifying them, and it has a suitable recognition for classifying texts
fourth research question	The combination of multilayer neural network and random graphs has brought the accuracy of diagnosis to 95% and the accuracy of diagnosis to 99%, which is a significant improvement.
fifth research question	By using random graphs, the amount of information to be processed has been reduced, and as a result, the output of the network has been improved.

[DOI: 10.61186/itrc.15.1.24]