ISSN: 2410-7727

TEXT INFORMATION CLASSIFICATION USING NEURAL NETWORKS

تصنيف المعلومات النصية باستخدام الشبكات العصبية

Dr.Hasan Abdullah Ahmed Al-Shaikh*

* *Department of information technology*
AL Andalus University
Sana'a, Yemen
Email:dr.hasan.alshaikh@gmail.com

ISSN: 2410-7727

ISSN: 2410-7727

TEXT INFORMATION CLASSIFICATION USING NEURAL NETWORKS

Dr.Hasan Abdullah Ahmed Al-Shaikh*

* *Department of information technology*
AL Andalus University
Sana'a, Yemen
Email:dr.hasan.alshaikh@gmail.com

ABSTRACT

This paper examines the theoretical underpinnings of machine-based text classification, a topic that has garnered significant interest recently. It outlines the primary stages and associated challenges in addressing these issues, presenting data derived from a straightforward algorithm for classifying text information. Key processes such as preliminary text filtering, feature vector formation, and the discussion encompasses the framework and training methodologies of a neural network. Performance outcomes are assessed utilizing the F-measure.

Comparative analysis is conducted on three different text corpora, investigating variations in preliminary filter parameters, hidden layer neuron counts, and network training durations. The proposed classifier model achieves an accuracy exceeding 80%, with the quality of training data playing a crucial role in this success.

The paper concludes by evaluating result quality and suggesting avenues for future research in this domain.

Keywords: semantic similarity, neural network, machine learning, text analysis, document classification.

ISSN: 2410-7727

تصنيف المعلومات النصية باستخدام الشبكات العصبية

د. حسن عبدالله احمد الشيخ*

*قسم تقنية المعلومات، جامعة الاندلس، صنعاء ، اليمن.

ملخص البحث:

تتناول هذه الورقة الأسس النظرية لتصنيف النص المعتمد على الآلة، وهو موضوع حظي باهتمام كبير في الآونة الأخيرة. وهو يوضح المراحل الأولية والتحديات المرتبطة بها في معالجة هذه المشكلات، ويقدم البيانات المستمدة من خوارزمية مباشرة لتصنيف المعلومات النصية. العمليات الأساسية مثل التصفية الأولية للنص، وتكوين ناقلات الميزات، والمناقشة تشمل إطار العمل ومنهجيات التدريب للشبكة العصبية.

يتم تقييم نتائج الأداء باستخدام مقياس F.

يتم إجراء تحليل مقارن على ثلاث مجموعات نصية مختلفة، لدراسة الاختلافات في معلمات التصفية الأولية، وأعداد الخلايا العصبية للطبقة المخفية، وفترات تدريب الشبكة. ويحقق نموذج المصنف المقترح دقة تتجاوز 80%، حيث تلعب جودة بيانات التدريب دوراً حاسماً في هذا النجاح.

وتختتم الورقة بتقييم جودة النتائج واقتراح سبل للبحث المستقبلي في هذا المجال.

الكلمات الافتتاحية: التشابه الدلالي، الشبكة العصبية، التعلم الآلي، تحليل النص، تصنيف الوثائق.

I. INTRODUCTION

The extensive volume of textual data in both public and private enterprise databases necessitates the utilization of robust analytical tools, particularly for the purpose of automatic categorization. A significant challenge in text classification tasks using machine learning is that textual data cannot always be precisely assigned to a specific class based solely on word frequency within a document. Text classification involves organizing unstructured text into predefined categories using Natural Language Processing (NLP). This process is also referred to as text categorization or text tagging.

Text data possesses an inherently sequential nature and high dimensionality, attributed to the extensive vocabulary involved in linguistic representation. Prior to engaging with Neural Networks, it is imperative to preprocess this data utilizing techniques such as tokenization, stemming or lemmatization, and vectorization.

The standard architecture of neural networks in Natural Language Processing (NLP) typically comprises several key components: an embedding layer that

ISSN: 2410-7727

transforms words into dense vectors; convolutional layers that apply filters to the embedded text; pooling layers, often max or average, which reduce dimensionality; fully connected layers that interpret the extracted features; and output layers dedicated to classification. Each element is crucial for comprehending contextual cues within textual data.

Algorithms for textual data classification facilitate the automated determination of a text's thematic category. This subset of automatic classification encompasses various tasks, including sentiment analysis, spam filtering, author identification, and the sorting of letters, messages, news articles, among other functions [10].

Training a neural network entails systematically adjusting the weight parameters through the iterative processes of forward and backward propagation.

Initially, during forward propagation, random weights are assigned to the inputs to generate an output. Subsequently, backward propagation is employed, wherein the margin of error between the predicted and actual outputs is calculated. This error is used to modify the weights in an attempt to minimize inaccuracies.

These cycles of forward and backward propagations continue until optimal weights are achieved for accurate predictions.

Within hidden layers of a neural network, activation functions play a critical role by summing weighted inputs and mapping them to appropriate outputs. Commonly used activation functions include linear, sigmoid, and hyperbolic tangent, among others.

And With the advent of open access to large amounts of textual information available in electronic form, there is a growing need to classify this information into different categories and to identify various patterns inherent in a certain group of textual data from of a certain sample. Despite the sharp increase in interest in problems of this kind in recent years, the development of new highly effective methods and means of classification is highly relevant [1]. The most important areas in natural language processing include thematic analysis of textual information. Thematic analysis allows you to divide text data into categories, for example, for quick classification to simplify the work of a person with these texts, for information systems search, as well as for some dialog Systems. This article describes the current problems in the field of textual information analysis, and reviews existing approaches and developments in this area.

ISSN: 2410-7727

II. RELATED WORK.

There are various studies on the subject of data classification using neural networks, shown in the following table1

. ***Table.1-Comparison between previous studies***

NO	Author	Algorithm	Dataset Taken	Results
1	Fouzi Harrag, Eyas Al-Qawasmah	Artificial neural network	Hadiths	SVD is 52.4s times faster than MLP NN, MLPNN gets best performance of 52%
2	Siwei Lai, Liheng Xu, Kang Liu, Jun Zhao	RCNN was proposed	20Newsgroups, Fudan Set, ACL Anthology Network,and Sentiment Treebank	Compared to traditional methods neural Networks performs well and comparing. The CNN, Recursive NN convolution based algorithms performs well
3	Jaromir Veber Mark HUGHES, Irene LI , Spyros	Information gain and Neural network	Reuters-22173 and OHSUMED collection	Train set accuracy was 95% and it is expected 90% accuracy on the real data
4	KOTOULAS and Toyotaro SUZUMURA	Word2vec	PubMed collection	Outperforms several natural

ISSN: 2410-7727

				language Processing algorithms by 15%
5	SaravananK and S. Sasithra	Dempster shafer theory of Evidence and neural networks	MEDLINE collection) AND MeSH thesaurus Newspage.com,	It is observed that the usage of demster shafer theory on the top Of neural network performs well
6	Taeho Jo	Neural text categorizer	20NewsGroups, and Reuter 21578	NTC is more practical than other classification techniques
7	E.N. Karuna, P.V. Sokolov	Convolutional neural network	Russian-language texts	The results show that bag-of-words based algorithms are easy to configure and they also have a high learning rate. But from the very structure of the algorithms it is clear that they have low accuracy when working with texts with complex semantics.
	Hasan. alshiakh	thematic analysis algorithm and neural networks	Reuters-21578, 20NewsGroups, myCorp	The results obtained in this article will become the basis for

ISSN: 2410-7727

				creating a cluster data analysis system, which will eliminate the need to manually label each text in a situation where there is a large corpus of texts that needs to be divided into topics.

From the table of previous studies and the current research, it is clear that each study was concerned with classifying a specific type of data and using different algorithms, while the study that I will conduct in this research paper is a classification of different data.

III. RESEARCH APPROACH

The general structure of the thematic algorithm analysis is shown in Fig. 1, detail machine learning algorithms for natural language processing problems are considered [2]. The main elements of the structure include: text database - source corpus of texts for training and testing the classifier,

ISSN: 2410-7727

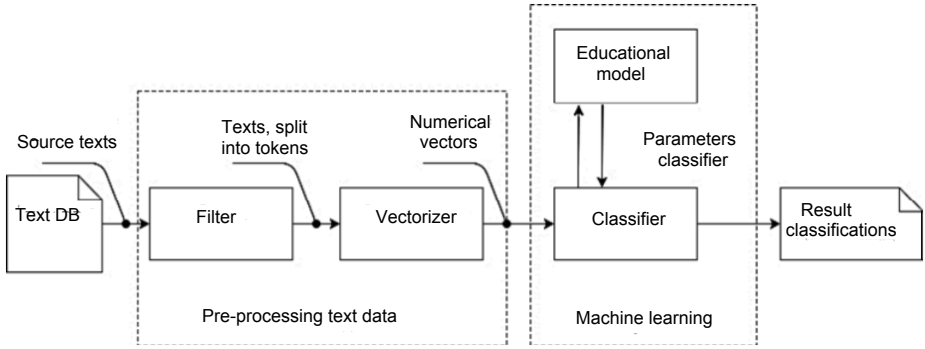

Fig. 1- General structure of thematic analysis algorithm

Filter: algorithm for removing non-letter characters, common words, reducing words to a single case, reducing to the base of a word.

Vectorizer: an algorithm for generating a numerical array with a set of features characteristic of each text.

Classifier: an algorithm for assigning a specific class to specific texts.

Training model: an algorithm for finding common patterns between input data and tuning the classifier to correctly divide the input data into appropriate classes [3]. A text database is a set of texts, when learning they should be are marked by topic. The whole group of texts it is divided into training and test samples. The article uses 3 corpora of texts, on which the classification algorithms are tested.

Collection Reuters-21578 contains about 20 thousand texts. In this set, the texts are unevenly distributed across topics. For experience Texts on the 10 most popular topics were selected.

myCorp Collection – a group of 500 texts on 10 different themes, collected by hand. The texts in this selection are quite easy to separate by topic, since the topics practically do not intersect with each other in meaning. The disadvantage of the collection is its small size compared to other samples.

The feature vector is compiled based on the frequency of word occurrence in the text. As part of the development of a text data classification program, an algorithm for their primary processing was implemented. The algorithm parses the text, when this ignores the service characters found in it. For the received words, it is necessary to perform the operation of finding the stem of the word (it will get rid of the situation when the same words, but written using different endings, are perceived as different words), for this a fairly common algorithm for finding the stem of the word is used - Porter stemmer [2].

Then the local dictionaries of each text are filled. At this stage, commonly used words in the English language are filtered out, using a set of 200 most popular words that are not associated with specific areas. This allows remove from consideration words whose presence does not contribute in any way to understanding the subject matter of the text. These include: conjunctions, prepositions, pronouns, function

ISSN: 2410-7727

words, and so on [4]. All other words end up in local dictionaries. After the analysis of all texts is completed formation of a global dictionary that contains all words from a cluster of texts. To further reduce the dimension of the dictionary, the following word filtering step is performed: words are removed whose total occurrence within the entire sample of texts is below the threshold value ε.

This type of filtering allows significantly reduce the vector dimension signs due to the removal of rare ones words The contribution of such words to the quality of work classifier is small, since the value of each of them inside each feature vector there will be extremely low and, as a result, uninformative. This will also allow you to remove words that, for various reasons, contain spelling errors. The filtering results can be seen in the table.2.

Tabel.2- Data filter results

Filtration parameters	Dimension of the feature vector for different collections of texts		
	Reuters	20NewsGroups	myCorp
Number of texts	7447	9931	500
Before filtering	44 821	53 692	26 382
$\varepsilon = 5$	6489	18248	7488
$\varepsilon = 10$	3830	11633	4782
$\varepsilon = 20$	2316	7583	3037
$\varepsilon = 30$	1776	5815	2335
$\varepsilon = 50$	1227	4128	1567
$\varepsilon = 100$	743	2572	827
$\varepsilon = 200$	431	1500	371
$\varepsilon = 400$	211	779	127

After receiving the text data divided into individual words, an algorithm for creating a feature vector for each text was implemented. The dimension of this vector directly depends on the size of the dictionary belonging to a given collection of texts. The vector of features consists of meanings, each of which allows you to characterize the frequency of occurrence of words in the text, according to the formula

ISSN: 2410-7727

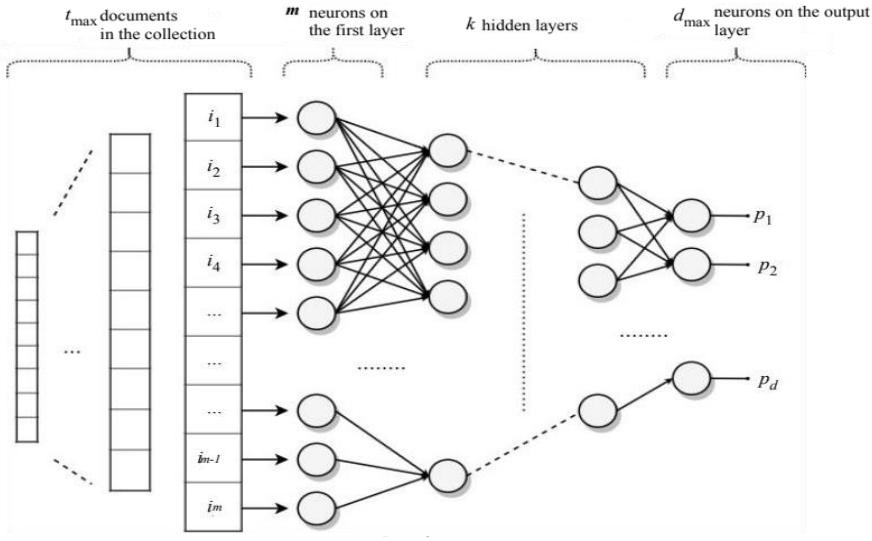

Fig. 2- Neural network

$$tf_i = \frac{m_{id}}{\sum_{j=1}^{i_{max}} m_{jd}} \quad ,$$

Where tf_i– Frequency of use of word i in document d; m_{id} – Number of words i in document; i_{max} – dictionary size;

$\sum_{j=1}^{i_{max}} m_{jd}$ – Number of words in the document d.

The neural network was created using the Neuroph library for the Java language [8]. Using individual objects from this library—neuron, layer, activation function, learning rules, and others we developed the neural network shown in Fig. 2.

The neural network is represented by a direct network distribution. The backpropagation method was used to train the network. Each of its nodes is represented by an artificial neuron, which serves as an analogue of the model natural neuron. Received at the entrance linear combination of all input signals passes through the sigmoid activation transfer function to send the signal to the neurons of the subsequent layer of the network [5].The size of the network input layer depends on the size dictionary of the corpus on which it occurs network training. The number of hidden layers k can be anything including 0, and the optimal value must be determined empirically. Size output layer depends on the number of topics in the example under consideration, this value is always equal to 10. The values on the output neurons are characterized by determine the probability of a text belonging to

ISSN: 2410-7727

each of the classes [9]. So, when determining whether a text belongs, the neuron with the highest output value is found - the number of this neuron will determine the class number.

The Reuters text collection contains an uneven distribution of texts by topic, which is why the standard assessment of the accuracy of the algorithm, depending on the ratio of correctly guessed topics to the total number of texts, is not suitable, since if the texts of the training sample are strongly biased towards one class, the classifier can accept adequate solutions only for this class. This leads to the fact that the classifier can produce a good overall accuracy rating, but the accuracy of identifying individual classes will be almost zero. For evaluation, it is necessary to use the F-measure, since this metric takes into account the average classification quality indicators for each class, and not their total classification accuracy. Calculating the F-measure to evaluate the solution to a classification problem with a number of classes greater than two requires constructing an error matrix W of size N × N, where N is the number of classes. The columns of this matrix are responsible for the number of the class to which document *d* belongs, the rows are responsible for the number of the class to which the classifier assigned this document. Then, when solving the classification problem for each document, there is an increase by one element in the element at the intersection of the required row and column - the resulting matrix allows you to clearly see the results of the classifier's work [5].

It is necessary to calculate the average accuracy $\overline{E}$ and recall $\overline{R}$ over the entire error matrix:

$$\overline{E} = \sum_{i=1}^{n} \frac{W_{i,i}}{\sum_{K=1}^{n} W_{i,k}} \qquad \overline{R} = \sum_{i=1}^{n} \frac{W_{i,i}}{\sum_{K=1}^{n} W_{k,i}}$$

Where i– number of the line for which the precision $\overline{E}$ is calculated, or the column number for which the completeness $\overline{R}$ of the error matrix is calculated; W – Error matrix; K – each row element when calculating the precision $\overline{E}$ or each column element when calculating the completeness $\overline{R}$;

The F-measure is calculated as the harmonic average between precision and recall:

$$F = 2\frac{\overline{E} \times \overline{R}}{\overline{E} \times \overline{R}}$$

ISSN: 2410-7727

IV. Results

First, it is necessary to determine the influence of the filter parameter ε on the quality of the classification algorithm. To do this, a series of experiments were carried out with different data sets on a network without hidden layers, while the neural network was represented only by input and output layers, the sizes of which depend on the dimension of the input vector and the number of topics, respectively. The results of the experiments are reflected in table. 3. According to the table:

$t_{training}$– Network training time; n – the accuracy of the classifier according to the evaluation method based on the ratio of the number of correctly guessed texts to the total number of texts in the corpus; F – classifier accuracy by F-measure; ε – filter parameter;

Table.3 - The results of the experiments

Algorithm parameters	Classification results for different document collections								
	Reuters-21578			20NewsGroups			myCorp		
ε	n	F	$t_{training}$	n	F	$t_{training}$	n	F	$t_{training}$
5	-	-	-	-	-	-	0.85	0.87	5:02
10	-	-	-	-	-	-	0.91	0.92	2:12
20	0.85	0.7	5:24	0.6	0.68	2:04:40	0.9	0.91	59 c
30	0.87	0.76	3:24	0.72	0.75	22:32	0.93	0.94	51 c
50	0.9	0.82	2:19	0.7	0.76	17:57	0.89	0.89	35 c
100	0.9	0.83	1:47	0.82	0.83	6:55	0.88	0.82	10 c
200	0.9	0.82	1:12	0.82	0.82	4:30	0.7	0.55	4 c
400	0.89	0.78	20 c	0.8	0.8	3:01	0.57	0.51	2 c

According to the table.3 shows that ε strongly influences both the speed of the network and the quality of classification. A decrease in ε causes the neural network to extract incorrect features corresponding to texts, which is why the network can continue to improve its performance on the training data, but perform worse on the test set due to the deterioration of the generalization qualities of the network [6].

The value of ε is selected depending on the size of the data sample. Thus, for the Reuters and 20NewsGroups text set, when cutting off all words whose occurrence is below 100 in the entire corpus, the highest level of performance is achieved for a neural network without intermediate hidden layers.

Next, a series of experiments is performed with a fixed value of ε and a variable number of neurons in the hidden layer table.4.

ISSN: 2410-7727

Table.4- The results of the experiments with network hidden layers

Algorithm parameters		Text collections								
		Reuters-21578			20NewsGroups			myCorp		
ε	Neurons in hidden layer	n	F	$t_{training}$	n	F	$t_{training}$	n	F	$t_{training}$
30	5	-	-	-	-	-	-	0.78	0.73	19
30	10	-	-	-	-	-	-	0.94	0.93	10
30	20	-	-	-	-	-	-	0.94	0.93	20
30	40	-	-	-	-	-	-	0.9	0.91	40
100	10	0.91	0.84	1:55	0.8	0.83	13:04	0.91	0.91	10
100	20	0.91	0.83	4:08	0.84	0.85	25:24	0.91	0.91	20
100	50	0.89	0.81	14:01	0.83	0.84	1:28:54	0.87	0.88	50

A comparison of changes in the accuracy of algorithms during network training with the best network parameters for each data sample is presented in Fig. 3.

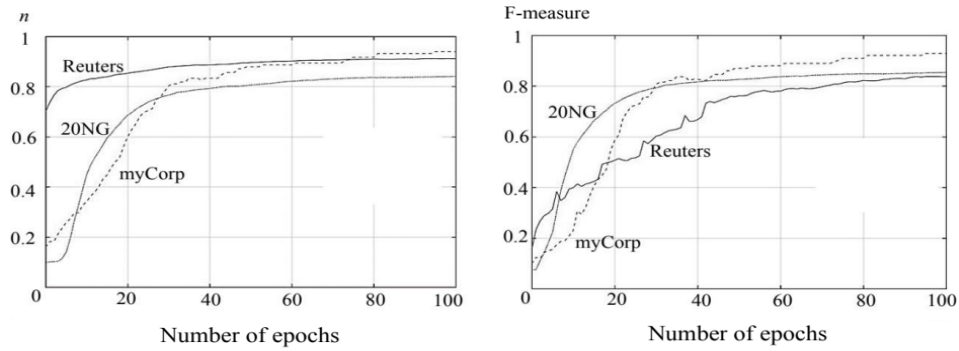

Fig. 3. Comparison of changes in the accuracy of algorithms during network training with the best network parameters for each data sample

The graphs show that the quality of the learning process strongly depends on the original data set. Thus, in a sample with an uneven distribution of topics, a fairly high frequency of guesses is observed at the very start of the algorithm, but at the same time, due to the low accuracy of determining any class to which a small number of documents belong, a low F-measure is observed. In this case, the best indicator turns out to be precisely in the smallest sample, and this is due to the fact that this corpus of texts has the least noisy source data and a uniform distribution of labeled data by topic.

For the Reuters-21578 text corpus, which is characterized by an uneven distribution of documents by topic with a strong bias of the training sample texts towards one class, the classification results are at the level of 91% accuracy in terms

ISSN: 2410-7727

of the ratio of correctly guessed texts to the total number of texts and 84% accuracy in terms of F-measure. The evaluations show that different classification results are observed for each class and that the best evaluation method is for the largest class of texts.

The 20NewsGroups corpus is the largest data sample; moreover, these documents contain a large amount of data that is not related to specific topics, which greatly complicates the classification task. The best classification scores are 84% accuracy and 85% F-metric, observed with $\varepsilon = 100$ and 20 neurons in the hidden layer.

Classification based on a prepared data corpus of 500 texts showed the greatest efficiency of the algorithm, providing an accuracy of more than 93% for two metrics with $\varepsilon = 30$ and 10 or 20 neurons in the hidden layer.

V. CONCLUSIONS

Based on the results obtained, the following conclusions can be drawn. The rare words filter, characterized by the value ε, has shown its effectiveness, but requires additional statistical research. The value of ε should depend on the size of the data corpus.

The quality of the training data plays a critical role in determining classification accuracy, and the quality turned out to be more important than the quantity of this data, since training on a sample of 500 texts showed better results than training on samples whose size was 14...20 times larger.

The proposed classifier model based on a neural network makes it possible to solve the classification problem has achieved an accuracy exceeding 80% across various evaluation methodologies. It has been determined that accuracy largely depends on the text corpus on which the network will be trained.

The results obtained in this article will become the basis for creating a cluster data analysis system, which will eliminate the need to manually label each text in a situation where there is a large corpus of texts that needs to be divided into topics.

ISSN: 2410-7727

References

1. Kowsari K, Jafari Meimandi K, Heidarysafa M, Mendu S, Barnes L and Brown D 2019 Text Classification Algorithms: A Survey Information 10 150.
2. Melnikov AV, Botov DS, Klenin JD. On usage of machine learning for natural language processing tasks as illustrated by educational content mining. Ontology of designing. 2017; 7(1): 34-47.
3. Stefan Wermter, Neural Network Agents for Learning Semantic Text Classification.
4. Gurmeet Kaur, Karan Bajaj News Classification using Neural Networks, 2016.
5. Siwei Lai, Liheng Xu, Kang Liu, Jun Zhao ,Recurrent convolution neural networks for text classification" in: proceedings of twenty ninth AAAI conference on artificial intelligence,2015.
6. Chaitanya Naik, Vallari Kothari, Zankhana Rana, "Document Classification using Neural Networks Based on Words", In: In- ternational Journal of Advanced Research in Computer Science,2015.
7. Siwei Lai, Liheng Xu, Kang Liu, Jun Zhao ,Recurrent convolution neural networks for text classification" in: proceedings of twenty ninth AAAI conference on artificial intelligence,2015.
8- E.N. Karuna, P.V. Sokolov. "Comparison of methods for automatic classification of Russian-language texts", Journal of Physics:Conference Series, 2021.
9. Akshat Tulsani, Jeh Patel, Preetham Kumar,Veena Mayya, Pavithra K.C., Geetha M.,Sulatha V. Bhandary, Sameena Pathan. "Anovel convolutional neural network for identification of retinal layers using sliced optical coherence tomography images",Healthcare Analytics, 2023.
10. Mingshi Zheng ,College of Computer and Information Sciences, Fujian Agriculture and Forestry University, Fuzhou, 350002, China, Research on text classification based on neural networks February 2024 ,Applied and Computational Engineering 41(1):282-303.