Scholar Publishing Group
Machine Learning Theory and Practice
https://doi.org/10.38007/ML.2020.010103
ISSN 2790-0983 Vol. 1, Issue 1: 22-32

Machine Learning Based News Text Classification

Tian Zhang[*]
Changchun Normal University, Changchun 130032, China
zhangtianzt2020@126.com
[*]*corresponding author*

Keywords: Machine Learning, News Text, Text Classification, Feature Extraction

Abstract: With the rapid development of the Internet era, the explosive growth of news data volume and the lack of effective management are gradually becoming serious problems, and it is increasingly difficult for readers to obtain valuable information quickly. How to quickly search for valuable information from the large amount of news text information is a meaningful task in text classification. Existing research methods still have some problems, such as directly combining the headline with the text, thus neglecting the importance of the headline, and the single model of classification, which leads to low classification results. For this reason, the main objective of this paper is to investigate the classification of news texts based on machine learning. This paper examines the current state of deep learning-based text classification and, in combination with the characteristics of news texts, chooses a machine learning-based text classification method to further explore and study news texts. Through the response time and core algorithm accuracy test of this system, the system better reflects the excellent performance of the system and meets the actual performance requirements of the system. The proposed news system can not only process system requests quickly, but also has excellent accuracy rate, which can better assist users to filter information and improve the user's experience of reading news.

1. Introduction

With the progress of the industrial internet, the internet has gradually turned into an indispensable platform and tool for people to exchange ideas and express their opinions. New media tools on the Internet have become an important tool for the Party, government and enterprises to strengthen public opinion guidance and promote social governance innovation [1-2]. Due to the complexity and diversity of online content, it needs to be classified to meet the needs of users for online opinion analysis [3-4].

In a related study, Mumtahina et al. proposed an integrated LIME BiLSTMBiLSTM to ensure classification accuracy and LIME to ensure transparency and interpretability of the model [5]. In this integrated model, the proposed model becomes easy to understand as LIME behaves similarly to the original model and explains the predictions. The performance of the model in terms of interpretability was measured using Kendall's tau correlation coefficient.Mohammadreza et al. aimed to address the lack of fake news datasets in Persian by introducing a new dataset crawled from different news agencies [6] and proposed two deep models based on the Bidirectional Encoder Representation by Transformer (BERT) model, which is a deep contextual pre-trained model for extracting valuable features. The results demonstrate the effectiveness of deep contextual embedding methods in the task of false news detection.

Based on the practical requirements of news systems, this paper provides system users with convenient news collection and persistent storage, news classification and summary results display, and other common functions, so that users can quickly read concise and concise news summaries under specific news categories and improve reading efficiency. As most news and information websites are currently divided into traditional categories, this affects the user's ability to dig into finer-grained news content that is of more interest to them; while news text summary details are not yet common in news websites.

2. Design Research

2.1. Key Issues in News Text Classification

Through the study of news text classification algorithms, it was found that traditional machine learning methods are prone to losing useful semantic feature information in the process of text representation, while using models such as Word2Vec and glove for text representation and then sharing text contextual semantic information by training neural network models can learn more vector representations as features, which is significantly better than traditional machine in terms of classification accuracy learning methods [7-8]. However, models such as Word2Vec cannot solve the problem of multiple meanings of words, especially in the face of the sparse features and context-dependent nature of news headlines, and there are still many semantic problems to be solved [9-10]. At present, the idea of deep neural network-based news text classification is to have deep networks automatically complete the extraction of features for efficient and accurate classification, and models such as TextCNN and LSTM are generally used [11-12]. textCNN can only capture local features due to its structural characteristics, and has weak extraction ability for long sequence dependencies. The feature extraction ability of LSTM and its variants is the opposite of LSTM and its variants have the opposite feature extraction capability to TextCNN. Therefore, the key to this problem lies in how to perform text representation and design network models to improve the accuracy of news text classification [13-14].

For this system, most of the classification problems of current news systems are implemented in news categories containing traditional broad categories such as politics, military, finance and economics [15-16]. And the news reported by the news media presents diversity and richness. Traditional news text classification can no longer meet the actual needs. Therefore, it is one of the key points of this paper that how to expand the traditional news categories make more fine-grained category classification and maintain a high accuracy rate at the same time with a small amount of manpower and time cost [17-18].

2.2. Text Classification Methods

In text classification, the principle of neural network-based classification methods is to perform vector representation of news text through natural language processing techniques, and automate feature extraction of text based on neural networks, and finally form an end-to-end text classification model through classifiers [19-20]. The following is a brief description of each algorithm.

Support vector machine (SVM): SVM is a classical classification algorithm for machine learning. Its principle is to maximize the class distance of samples by using the point-to-hyperplane distance calculation to achieve sample classification.

Convolutional Neural Network (CNN): The principle is to use the convolution and pooling process in the convolutional neural network structure to extract the key words and phrases from the text, and then combine it with a classifier to achieve classification.

BiLSTM (Long Short-Term Memory): The principle of BiLSTM is to use the "gating" mechanism of the Long Short-Term Memory network to control the level of memory of text sequences, and the bi-directional nature of the network allows it to accurately extract contextually important text features and then use the Softmax classifier to classify them.

BiLSTM-Attention (Long Short-Term Memory with Attention): The principle is to add an attention mechanism to the Bi-LSTM network to enhance the representation of important features through the attention mechanism.

FastText: The principle is to train logistic regression using n-gram vectors of sequence context, and then implement a Softmax classifier for classification.

Bidirectional Encoder Representation from Transformer (Bert): advanced language model, which is based on the principle of a structure made by stacking multiple layers of Transformer structures with powerful language representation [21-22]. The advantages and disadvantages of each classification method are shown in Table 1.

Table 1. Comparison of the advantages and disadvantages of each classification method

Methods	Advantages	Disadvantages
SVM	High classification accuracy for small samples and good generalisation ability of the model.	Difficult to train on large scale data, cannot directly support multi-classification.
CNN	Effective in extracting global features of text, with high classification accuracy.	Supervised learning approach, which requires large data volumes to train the model.
BiLSTM	Efficient extraction of text contextual features, high classification accuracy and good generalisation ability.	Supervised learning method also requires large data size to train the model.
FastText	High classification accuracy, robustness and speed.	Does not support multi-label classification and requires high quality training samples.
Bert	Very high classification accuracy and strong model generalisation.	Good equipment and hardware are required to train large-scale data.

2.3. Machine Learning Methods

(1) Parsimonious Bayesian

The Parsimonious Bayesian (NB) algorithm is simple to understand and efficient to learn. It calculates the posterior probability that a Y variable belongs to a category by means of a number of prior probabilities. Although the plain Bayesian model is relatively simple, it works well in many applications such as sentiment classification. However, in intention recognition tasks, plain Bayesian uses the product of the probabilities of the words as the basis for classification discrimination, resulting in an inability to focus on local features. The following equation shows the principle of plain Bayesian formulation.

$$P(Y \mid X) = \frac{P(X \mid Y)}{P(X)} = \frac{P(X \mid Y)P(Y)}{P(X)} \tag{1}$$

(2) Support vector machine

Support vector machines (SVMs) assume that the binary classification problem is linearly separable, partitioning an m-dimensional vector space with an m-1 dimensional plane such that the two classes of data are furthest apart from the partition plane. A schematic representation of the classification principle of a support vector machine is shown in Figure 1.

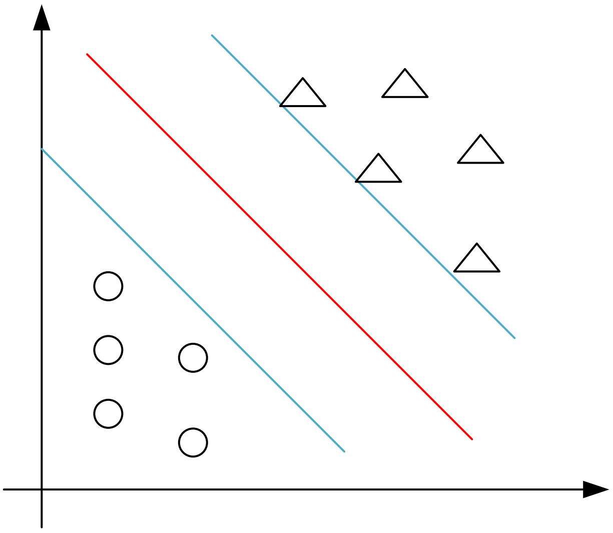

Figure 1. Schematic diagram of a support vector machine

3. Experimental Study

3.1. Feature Extraction

Feature extraction of the data is an essential step when modelling using machine learning algorithms. Feature extraction is a very important part of the data pre-processing process, and feature extraction can also lead to a reduction in the dimensionality of the data. It is important to note that important information must not be lost during the feature extraction process otherwise the training process of the machine learning algorithm will be seriously affected, resulting in the algorithm not learning more information.

(1) Cardinality test

The basic idea is to calculate the correlation between two variables and then do correlation on the feature attributes. In text analysis, the cardinality test is used for feature selection to calculate the correlation between a word and a category. Typically the chi-square value 2x (t,c) is calculated as

$$x^2(t,c)\frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)'} \tag{2}$$

In the above equation, c denotes a categorical category; t denotes a word feature value; N=A+B+C+D denotes the total training text set; A denotes the number of texts in a category that contain a t feature value; B denotes the number of texts in other categories that contain a t feature value; C denotes the number of texts in a category that do not contain a t feature value; and D denotes the number of texts in other categories that do not contain a t feature value.

In equation (2) N, A+B and C+D are all fixed values then it can be simplified as:

$$x^2(t,c)=\frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)'} \tag{3}$$

The chi-square test can be useful for extracting features, but it also has the drawback that it only counts whether a word occurs in a document, and does not indicate how often a word occurs.

(2) Reciprocal information

The usual formula for the mutual information I (w,c) is

$$I(w,c)=\log\frac{\frac{A}{N}}{\frac{A+B}{N}\frac{A+C}{N}}=\log\frac{A\cdot N}{(A+C)\cdot(A+B)'} \tag{4}$$

where I (w,c) denotes relevance; N is the number of training texts; A denotes the number of texts that feature word w is divided into category c; B denotes the number of texts that feature word w is divided into that are not in category c; C denotes the number of texts that do not belong to feature word w is divided into category c; D denotes the number of texts that do not belong to feature word w is divided into that are not in category c.

(3) Information gain

Information gain is widely used in machine learning algorithms, such as the calculation of the root node of a decision tree, and it is also one of the most commonly used feature extraction methods. The basic idea of information gain is also to calculate the relationship between feature attributes and categories. Usually the information gain IG (W,C) is calculated by the formula

$$IG(W,C)=H(\mathrm{C})-H(C\,|\,W) \tag{5}$$

$$H(\mathrm{C})=-\sum_{i=1}^{n}p_i\log p_i \tag{6}$$

$$H(C\,|\,W)=\sum_{i=1}^{n}P(W=w_i)H(C\,|\,W=w_i) \tag{7}$$

Where IG(W,C) denotes the information gain value of attribute W feature value under category C; H(C) denotes the information entropy of category C and H(C|W) denotes the conditional entropy of attribute feature word W in category C.

3.2. Classification Performance Evaluation

When machine learning classification algorithms are used to build classifiers, different classifiers are learned using a "training dataset" that has been pre-processed with data. To evaluate the performance of the classifier, a portion of the pre-processed 'test dataset' is then used to test the generalisation of the classifier. When using machine learning algorithms, metrics such as accuracy, recall and F1 scores are generally used. For ease of understanding, a confusion matrix of classification results is created in Table 2.

Table 2. Confusion matrix of classification results

Confusion matrix	Prediction Positive	Predicted Inverse
Prediction Positive	TP	FP
Predicted Inverse	FN	TN

Where TP + FN + FP + TN = total number of samples.

The accuracy rate P is defined as the probability that the number of news items classified into the category is greater than the sum of the number of news items classified into the category plus the number of news items actually not in the category but classified into the category, i.e. the number of news items correctly classified into the category. The specific formula is as follows:

$$P = \frac{TP}{TP + FP} \tag{8}$$

Recall R is defined as the probability that the number of news items classified into that category is greater than the sum of the number of news items classified into that category plus the number of news items that are actually in that category and are classified into other categories. The specific formula is as follows.

$$P = \frac{TP}{TP + FN} \tag{9}$$

To combine the performance of the classifiers, the F1 score is obtained by combining the accuracy and recall and using this as an evaluation metric for another classifier. The specific equation is as follows.

$$P = \frac{2RP}{P + R} = \frac{2TP}{2TP + FN + FP} \tag{10}$$

The above three evaluation metrics can better evaluate the performance of the classifier, but in practical applications, it is not enough to consider only the above three metrics, but usually the algorithm complexity is also considered as the last evaluation metric. The complexity of the algorithm is usually measured in time, and when the time complexity of the algorithm is too high, it leads to high time costs when performing training and testing.

3.3. Experimental Analysis

(1) Experimental environment and data set

The experimental environment parameters were set as shown in Table 3.

Table 3. Experimental environment and parameter settings

Software environment	Interpreters	Python3.6.8
	Operating systems	Ubuntu16.04
	Integrated Development Environment	Pycharm
	Deep learning frameworks	Tensorflow1.12
Hardware environment	CPU	i7-7700k CPU
	GPU	Nvidia GeForce GTX 1080
	Memory	16GB

One is the nlpcc2017 dataset, which is a Chinese news headline dataset containing 18 news types, including military, current affairs and science and technology, with 228,000 samples. The other is the thucnews dataset, which is a Chinese news body dataset containing 10 news types with 65,000 samples.

(2) Experimental scheme for news text classification

(1) Validation of the classification model with data augmentation strategy. This experiment firstly determines the optimal parameters of the proposed data enhancement strategy; and then verifies the effectiveness of the proposed data enhancement strategy through comparison experiments using the proposed classification model as the benchmark.

(2) Validation of the classification model incorporating semantic enhancement. The experiments will be based on the current typical classification models for comparison, to verify the validity of the proposed semantic enhancement-based classification model. At the same time, the model performance will be further validated by combining different experimental datasets with control variables.

4. Experiment Analysis

4.1. System Response Time Testing

The performance tests in the system are designed to obtain data on performance indicators under certain conditions. The test points are the various functional modules of the system (including basic and advanced functions). The test method was used to simulate 100 access operations by 10 users. Table 4 records the results of the system performance tests for each metric.

Table 4. System performance test results of each index test record table

Test Items	Login	Register	Message Modification	News Categories	News summary
Total execution time (ms)	943	972	1552	12113	17031
Throughput QPS (ms)	106	124	39	8	5
Minimum response time(ms)	41	56	65	231	355
Maximum response time (ms)	341	368	1677	4396	4657
Average response time (ms)	92	103	182	1067	1582

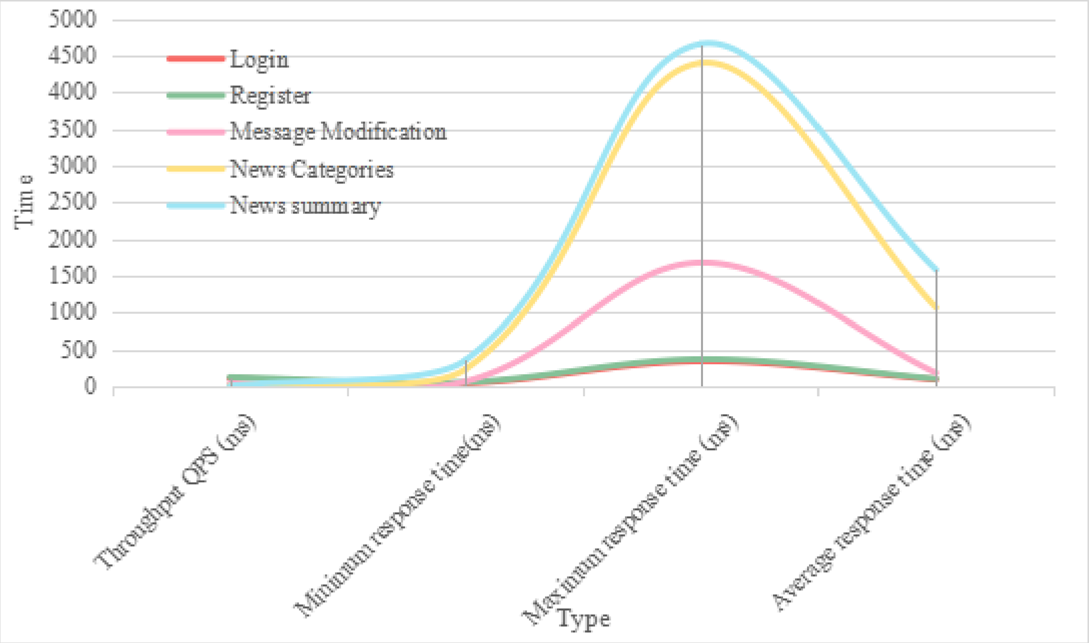

Figure 2. Analysis of test results for each indicator of system performance test

As can be seen from the records in Figure 2, the overall performance of the system is good, with high throughput and short average response time for the basic functions (login, registration, information modification); the throughput and average response time for the advanced functions (news classification, news summary) are more substantial and basically meet the system performance requirements.

4.2. System Core Algorithm Accuracy Verification

To evaluate the accuracy of the core algorithm functions of the system, a test dataset was annotated with some of the data provided by a research institute and the collected news data. This dataset contains both news classification and news summary data. Similarly, the news classification algorithm was tested using the P, R and F1-Score metrics, and the news summary algorithm was tested using the Rouge-1, Rouge-2 and Rouge-L metrics, with the results shown in Table 5.

Table 5. Accuracy testing of core system algorithms

Test item		Target sample categories (pcs)	P(%)	R(%)	F1-Score(%)
News Categories (Headlines)	Test1	50	92.03	92.45	91.96
	Test2	100	92.85	92.33	92.56
News Categories (Text)	Test1	50	95.63	96.15	95.66
	Test2	100	95.72	95.39	96.04
Test Items		Target sample size (pcs)	Rouge-1	Rouge-2	Rouge-L
News Summary	Test1	100	0.5602	0.3430	0.5325
	Test2	150	0.5423	0.3333	0.5285

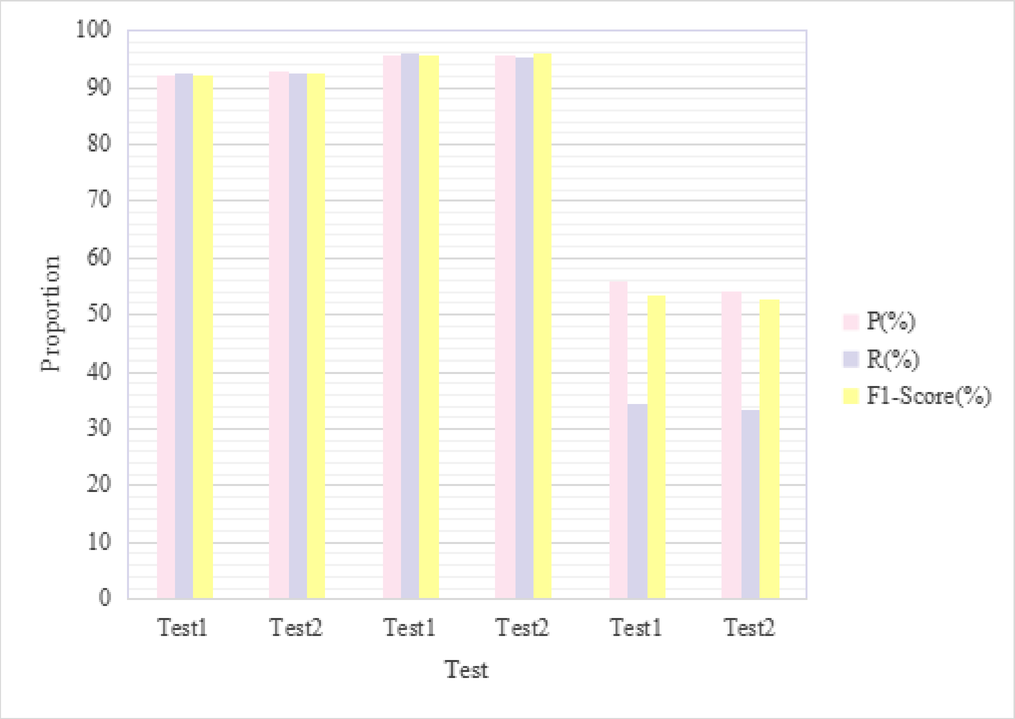

Figure 3. System core algorithm accuracy test analysis diagram

As can be seen from Figure 3, through the response time and core algorithm accuracy test of this system, the system better reflects the excellent performance of the system and meets the actual performance requirements of the system. The news system proposed in this paper can not only process system requests quickly, but also has excellent accuracy rate, which can better assist users to filter information and improve their experience of reading news.

5. Conclusion

The times are progressing and internet technology is developing. While reaping the benefits of the information age, there are also many pressing issues that need to be addressed. For example, in the face of the massive amount of news text information, how to obtain valuable information in an efficient manner has become the focus of research. This paper first discusses the current state of research on text classification based on traditional machine learning, and the limitations in the field of news text classification. In addition, the current status of deep learning-based text classification is investigated, and in combination with the characteristics of news texts, this paper chooses a machine learning-based text classification method to further explore and study news texts. The research on news text classification is based on the following highlights: Firstly, although deep learning is effective in text classification, it is less used in the field of news text classification, and most of the research has not been conducted for the special format of news text. Second, it is found in the study that the singularity of text classification models also affects the classification results.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Mete Eminagaoglu: A new similarity measure for vector space models in text classification and information retrieval. J. Inf. Sci. 48(4): 463-476 (2020). https://doi.org/10.1177/0165551520968055

[2] Mumtahina Ahmed, Mohammad Shahadat Hossain, Raihan Ul Islam, Karl Andersson: Explainable Text Classification Model for COVID-19 Fake News Detection. J. Internet Serv. Inf. Secur. 12(2): 51-69 (2020).

[3] Aytug Onan: Bidirectional convolutional recurrent neural network architecture with group-wise enhancement mechanism for text sentiment classification. J. King Saud Univ. Comput. Inf. Sci. 34(5): 2098-2117 (2020).

[4] Dede Rohidin, Noor Azah Samsudin, Mustafa Mat Deris: Association rules of fuzzy soft set based classification for text classification problem. J. King Saud Univ. Comput. Inf. Sci. 34(3): 801-812 (2020). https://doi.org/10.1016/j.jksuci.2020.03.014

[5] Mumtahina Ahmed, Mohammad Shahadat Hossain, Raihan Ul Islam, Karl Andersson: Explainable Text Classification Model for COVID-19 Fake News Detection. J. Internet Serv. Inf. Secur. 12(2): 51-69 (2020).

[6] Mohammadreza Samadi, Maryam Mousavian, Saeedeh Momtazi: Persian Fake News Detection: Neural Representation and Classification at Word and Text Levels. ACM Trans. Asian Low Resour. Lang. Inf. Process. 21(1): 10:1-10:11 (2020). https://doi.org/10.1145/3472620

[7] Vitor Garcia dos Santos, Ivandr éParaboni: Myers-Briggs personality classification from social media text using pre-trained language models. J. Univers. Comput. Sci. 28(4): 378-395 (2020). https://doi.org/10.3897/jucs.70941

[8] Himashi Rathnayake, Janani Sumanapala, Raveesha Rukshani, Surangika Ranathunga: Adapter-based fine-tuning of pre-trained multilingual language models for code-mixed and code-switched text classification. Knowl. Inf. Syst. 64(7): 1937-1966 (2020).

[9] Prabhat Dansena, Soumen Bag, Rajarshi Pal: Pen ink discrimination in handwritten documents using statistical and motif texture analysis: A classification based approach. Multim. Tools Appl. 81(21): 30881-30909 (2020).

[10] Asad Masood Khattak, Muhammad Zubair Asghar, Hassan Ali Khalid, Hussain Ahmad: Emotion classification in poetry text using deep neural network. Multim. Tools Appl. 81(18): 26223-26244 (2020).

[11] Rajib Ghosh: A recurrent neural network based deep learning model for text and non-text stroke classification in online handwritten Devanagari document. Multim. Tools Appl. 81(17): 24245-24263 (2020).

[12] Ngoc Lethikim, Thao Nguyen-Trang, Tai Vovan: A new image classification method using interval texture feature and improved Bayesian classifier. Multim. Tools Appl. 81(25): 36473-36488 (2020).

[13] Abadhan Ranganath, Manas Ranjan Senapati, Pradip Kumar Sahu: A novel pixel range calculation technique for texture classification. Multim. Tools Appl. 81(13): 17639-17667 (2020).

[14] Jeow Li Huan, Arif Ahmed Sekh, Chai Quek, Dilip K. Prasad: Emotionally charged text classification with deep learning and sentiment semantic. Neural Comput. Appl. 34(3): 2341-2351 (2020).

[15] Hozayfa El Rifai, Leen Al Qadi, Ashraf Elnagar: Arabic text classification: the need for multi-labeling systems. Neural Comput. Appl. 34(2): 1135-1159 (2020).

[16] Gilles Jacobs, Cynthia Van Hee, Véronique Hoste: Automatic classification of participant roles in cyberbullying: Can we detect victims, bullies, and bystanders in social media text? Nat. Lang. Eng. 28(2): 141-166 (2020). https://doi.org/10.1017/S135132492000056X

[17] Lydia Binti Abdul Hamid, Anis Salwa Mohd Khairuddin, Uswah Khairuddin, Nenny Ruthfalydia Rosli, Norrima Mokhtar: Texture image classification using improved image enhancement and adaptive SVM. Signal Image Video Process. 16(6): 1587-1594 (2020).

[18] Soulib Ghosh, Khalid Hassan Sheikh, Hussain Ali Khan, Ankur Manna, Showmik Bhowmik, Ram Sarkar: Application of texture-based features for text non-text classification in printed document images with novel feature selection algorithm. Soft Comput. 26(2): 891-909 (2020).

[19] Amir Kenarang, Mehrdad Farahani, Mohammad Manthouri: BiGRU attention capsule neural network for persian text classification. J. Ambient Intell. Humaniz. Comput. 13(8): 3923-3933 (2020).

[20] Dangguo Shao, Chengyao Li, Chusheng Huang, Qing An, Yan Xiang, Junjun Guo, Jianfeng He: The short texts classification based on neural network topic model. J. Intell. Fuzzy Syst. 42(3): 2143-2155 (2020). https://doi.org/10.3233/JIFS-211471

[21] Kushagri Tandon, Niladri Chatterjee: Multi-label text classification with an ensemble feature space. J. Intell. Fuzzy Syst. 42(5): 4425-4436 (2020). https://doi.org/10.3233/JIFS-219232

[22] N. Venkata Sailaja, L. Padma Sree, Nimmala Mangathayaru: Statistically Empirical Integrated Approach for Knowledge Refined Text Classification. J. Inf. Knowl. Manag. 21(2): 2250027:1-2250027:21 (2020). https://doi.org/10.1142/S0219649222500277