Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 167 (2020) 2124–2131

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Computational Intelligence and Data Science (ICCIDS 2019)

Design of Kids-specific URL Classifier using Recurrent Convolutional Neural Network

R. Rajalakshmi[a,*], Hans Tiwari[b], Jay Patel[b], Ankit Kumar[b], Karthik.R[b]

[a]*School of Computing Science and Engineering, Vellore Institute of Technology, Chennai*
[b]*School of Electronics Engineering, Vellore Institute of Technology, Chennai*

Abstract

The use of digital devices has increased exponentially in the last decade. Especially, young children spend most of their time surfing for various reasons such as homework. assignments and projects etc. Parental Control is highly important for monitoring the browsing behavior of children. Though several content-based web page classification approaches are available, it requires the entire web page contents for classification purpose. This leads to wastage of bandwidth due to unnecessary downloads. The exponential growth of internet demands URL based classifiers to adapt to the dynamic web, and to make swift decisions on the fly. To address this problem, a deep learning based approach has been proposed in this research that can extract the features only from the URL of a web page. To learn the patterns for Kids-specific web sites automatically, Convolutional Neural Network (CNN) is combined with Bidirectional Gated Recurrent Unit (BGRU) to extract rich context aware features as well as to preserve the sequence information in the URL. By conducting various experiments on the benchmark collection Open Directory Project (ODP), it is shown that an accuracy of 82.04% can be achieved.

Keywords: Kids; URL Classification; Deep learning; RCNN; ODP dataset

1. Introduction

The browsing behaviour of children needs to be monitored in this digital world as the content available in web is not reliable. In general, children used to read the articles from web for doing their homework, school projects etc. They also have the habit of reading stories online and watch cartoons or movies. As the kids related content of the web page is highly dynamic in nature, the traditional method of content based classification is not suitable. A survey on information retrieval system for children has been studied by Tatiana [2]. They outlined the search query used by children, search strategy and the navigation style. The need for a separate page ranking algorithm for children

* Corresponding author.
E-mail address: rajalakshmi.r@vit.ac.in

10.1016/j.procs.2020.03.260

web pages has been discussed by these authors. Gyllstrom [3] suggested a link based algorithm to rank the child related web pages according to the relevance. Zhang et ál [1] analysed the role of URLs in finding the objectionable content, but in their approach, they have combined the content and the URLs for classifying the web page. But content based classification is not suitable, as the web content changes dynamically. So the only way to monitor the browsing behaviour of children is to make use of the links that they visit during their browsing time. To address this issue, a simple URL based classifier is suggested. The problem of classifying a web page as Kids related or not, is studied in detail by combining the advantages of Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN).

In this research, we have proposed a deep learning method to classify Kids-specific web page by extracting features from URLs alone. For URL classification, token based features were suggested in literature [5] [12]. Considering URL features alone has the issues like misspelled tokens, clueless / ambiguous terms. The handcrafted features like n-grams were derived from URLs and the relationship among the terms were not considered. In many cases, the tokens may be misspelled for making the URL catchy, which makes classification difficult. So, we have integrated an auto spelling correction method in this proposed system. The URL is a sequence of tokens that are delimited by some characters like :, /, ., - etc. Also it may contain terms that are clueless to guess the topic it talks about. Because some common terms may be ambiguous, for which different meanings are possible based on the terms associated with it. For example, a word (token) present in the URL may be 'driver', the meaning of this token is fully dependent on the other terms that are surrounding it. It may denote a 'device driver' or a 'car driver' depending on the context. In traditional methods of URL classification, only the terms present in the URL are taken into account in the form of token features or n-gram features [5, 10, 4, 15] and the context of the terms are not considered. In the proposed work, in order to obtain context of the tokens, we have used word embedding. The URL contains a sequence of inter-related tokens and hence we applied a Recurrent Neural Network and tried to make use of sequence information also for classification of web pages. Instead of relying on hand crafted features, thanks to Convolutional Neural Network, we have attempted to learn the rich features automatically from the URL.

The proposed method addresses all the above issues viz. automatic extraction of suitable URL features with necessary spelling correction and aslo preserves the context and sequence information in the URL. We have conducted various experiments on the bench mark collection Open Directory Project (ODP) dataset by considering 90,560 URLs and studied the effectiveness of various deep learning architectures. From the experimental results, we found that Bidirectional Gated Recurrent Unit (BGRU) with Convolutional Neural Network is the suitable method for classifying the web pages as Kids-specific or not, and we have achieved an accuracy of 82.04 % in this approach, which is a significant improvement over the exisitng URL based classification methods.

The paper is organized as follows: In Section 2, various approaches that applied machine learning techniques for web page classification have been discussed. The advantages of deep learning techniques such as Convolutional Neural Network and Recurrent Neural Network for various text classification problems were also presented. The proposed Recurrent Convloution Neural Network architecture for classifying the Kids relevant URLs has been elaborated in Section 3. The details about the experimental study and the result discussion are presented in Section 4. The comparitiive analysis is detailed in Section 5 followed by the concluding remarks in Section 6.

2. Related Works

The problem of web page classification is an important topic of research, as it has many applications like topic categorization, recommendation systems and focussed crawlers. Various approaches have been suggested in the literature to perform this task by considering various features viz., content based features, on-page features and link based features. But the content based approaches are not preferred often, as fetching the entire web page contents and processing those large volume of data requires high computational power, storage and also takes much time. To overcome these issues, Kan [11] has proposed an URL based approach, thereby avoiding the need for downloading the contents. To derive the useful features from the URL. they have applied segmentionan / expansion approach. They have conducted experiments on a small WebKB dataset.

Baykan et . al [5] have performed a detailed analysis of URL classification and considered various features such as tokens and all n-grams (n = 4 to 8). They have shown the importance of n-gram features with out applying any feature selection method, but their method depends on the dimensionality of the training set and not suitable for large scale data. By constructing a heuristic based dictionary from the token derived from URLs, Rajalakshmi et ál [12]

have tried to find the category of the web site. In this approach, the term frequency of the tokens were taken into account and based on the partial or complete match with the heuristics dictionary, tokens in the URL were given different weightage. But, this approach mainly depends on the dictionary and the size of this dictionary increases when the number of URLs is large, thereby imposing a challenge. In [14], an SVM based method was proposed by extracting only 4-grams from the URLs, health related domains were classified. This work was extended in [15] and an automatic learning of URL features was proposed. In this approach, instead of considering 4-grams alone, all the n-grams (n = 3,4,5,6,7,8) and token features were extracted to determine the category of the web page. A dataset independent dictionary construction method has been suggested. By conducting various experiments on ODP dataset and WebKB dataset, they have shown the effect of feature weighting methods. For filtering the web pages based on the classification score, a rejection framework has been included in the Naive Bayes URL based classifier. In this work, a method to embed the term goodness into the Naive Bayes classifier was suggested. The improtance of n-gram features have been analysed in [4] also.

Recurrent Neural Network (RNN) model analyzes a text, word by word and stores the semantics of all the previous text in a fixed-sized hidden layer. Contextual information capturing is the major advantage of this model. However, one of the drawbacks of it is the presence of bias which reduces the overall effectiveness when it is capturing the semantics of an entire document, the reason being the occurrence of key components anywhere in the document instead at only the end. In the URL also, the above problem exists as the topic / class of the web page can be predicted in a better way, if the entire sequence is considered independent of its position in the URL. Convolutional Neural Network (CNN), an unbiased model, was introduced to NLP to tackle the above mentioned issue. This is because it is able to determine discriminative phrases in an URL with a max-pooling layer. However, while using simple convolutional kernels as being used in a number of previous studies, it is difficult to determine the window size. There are a few issues with the approach, the main being that if the window size is smaller, we may lose important information, whereas if the window size is bigger, we may end up with an unnecessarily large parameter space, which results in difficulty in training. Zhou [18] has proposed attention based approach for classifying the short text in Chinese language. Yang [16] proposed a method for classifying the documents using hierarchical attention based models. For the purpose of URL classification, in order to overcome the limitations mentioned in the above models, Recurrent Convolutional Neural Network (RCNN) has been proposed.

All the above mentioned URL based methods try to select the suitable features by applying the traditional machine learning algorithms and considered the features from the URLs alone. The context of the terms in URL is not explored much and the power of deep learning techniques has not been utilized. In identifying the malicious domain from DGA generated domain names [17] , character level convolutional neural network has been applied and the features learnt from CNN were used to train the machine learning models such as Naive Bayes, XGB and SVM classifiers. This transfer learning approach is found to be better than the other methods. But, the sequence information was not considered in this approach.

In the proposed approach, word level features have been considered from the URLs and the combination of Recurrent Neural Network and Convolutional Neural Network has been employed to derive the context information and also to learn the features from URLs.

3. Proposed Methodolgy

The objective of the proposed work is to determine the topic of a web page as Kids specific or not. As the content based approaches are not suitable for this task, and due to the dynamic nature of web, URL based classification is performed.

An URL consists of sequence of tokens. For example, consider the URL, $http : //www.free - for - kids.com/printable - word - search - puzzles.html$. In this example URL, free, for, kids, com, printable, word, search, puzzles, html are said to be tokens. The input URLs are tokenized and then stop words, special characters are removed followed by Porter stemming. In general, the tokens in the URL need not be meaningful, so auto spell check feature is included. These pre-processed tokens are represented as a vector by adding an embedding layer. To learn the appropriate and rich features, Convolutional Neural Network is employed. Then, a Bidirectional GRU is used to obtain the sequence information in both the directions from the URL and a suitable deep learning model is built. This trained model is then used for predicting the category of an unseen URL as a Kids specific website or not.

The proposed system consists of the following stages viz.,

1. Input Layer
2. Embedding layer
3. Recurrent Convolutional Neural Network
4. Output layer

3.1. Input Layer

To make the URL suitable for input layer, the first step is to perform pre-processing. It includes the following steps: removing the special characters, converting all the alphabetical characters to lower case, removing the stop words using porter stemmer algorithm. Then tokens are separated. However, some tokens may be misspelled or shortened. To obtain the correct spelling for each token, we have added an auto spell correction feature using Symspell library.

3.2. Embedding Layer

As the preprocessed URL is of varying size, padding is performed to get a fixed length input. Instead of considering the token features alone, the context of the tokens need to be included in the URL representation. Global Vector for Word Representation (GloVe) [6] is an unsupervised learning algorithm used for representing words as vectors and helps in keeping similar words closer to each other. We have used this pre-trained word vector, which contains 6B tokens, 400K vocab (uncased), and 300d vectors for each word. We used Glove to get the word vector for the tokens in every URL after pre-processing.

3.3. Recurrent Convolutional Neural Network

In this section, we have presented the significance of Recurrent Convolutional Neural Network based architecture for URL classification.

3.3.1. Convolutional Neural Network

The application of Convolutional Neural Network (CNN) for the text classification problem has been increased in recent days, because of the significant performance of it. For this URL based classification, word based convolutions are applied. We have developed the model, by using CNN alone with pre-trained embedding from GloVe. The advantage of applying CNN for this task is that, a good representation of the input URL can be learnt automatically by the convolutional filters without providing the whole collection of tokens. CNNs are also found to be fast and effective. In this work, the width of kernel is fixed to 300, as we have used the 300 dimensional vector in GloVe and the length of the kernel is 3, i.e. we have restricted to 3 tokens at a time. So, these convolutional filters capture the features that are similar to word n-grams, but it is done in a more compact way. These filters are applied on a sequence of tokens to produce the feature map and the max pooling was applied on these feature maps.

In the CNN based Model, even though CNN learns the rich token features from the URL, the context information is not captured. It extracts the most informative tokens and considers their activations [7]. For identifying the Kids related web page, the entire sequence of tokens present in URL may be considered, as some tokens in the initial parts of the URL may be related to other tokens in subsequent parts. To achieve this, RNN based approach is proposed in this paper.

3.3.2. Recurrent Neural Network

By applying a Recurrent Neural Network (RNN) based deep learning architecture, the sequence information present in an URL could be obtained. By looking at an URL token, the RNN tries to derive the relation from the previous tokens in that URL. However, RNN will not derive contexts from the far behind tokens, for a lengthy URLs. To overcome this issue, Gated Recurrent Unit (GRU) has been widely used. In GRU based model, the amount of information to be passed to next state can be controlled. Here, an update gate is used to decide how much information is to be passed to the next node and a reset gate is used to decide how much information should be discarded.

Even though, the above mentioned GRU based model is able to capture the context information, it is restricted to forward direction alone. In an URL, the tokens that appear in future may depend on the previous tokens also. In order to capture the dependency and information in both the directions, we have used Bidirectional Gated Recurrent Unit (BGRU). In this BGRU model, the context information is captured in both the directions.

3.3.3. Bidirectional Gated Recurrent Unit with CNN

As discussed in Section 3.3.1 and Section 3.3.2, CNN learns rich token features from the URL and BGRU is used to capture context information in both directions of the URL. This motivated us to combine the advantages of CNN with BGRU architecture thereby important token features along with their context information can be derived from an URL. By this approach, we try to find the dependency between tokens present in different positions. As discussed in Section 1, the token driver may occur in the initial part of the URL, whereas the token device may appear at other part of the URL. In this case, the update gate in GRU is used to carry this information back and forth, whereas the reset gate is helpful to discard the remaining tokens that are not directly contributing to this.

3.4. Output Layer

The final output layer is the dense layer with the softmax activation function, which is used to decide whether the given URL is Kids related or not.

In this way, the URL based classification is performed using Bidirectional RCNN that combines the Recurrent Neural Network with Convolutional Neural Network.

4. Experiments and Results

The experimental set up and description of the datasets are presented below along with the discussion of obtained results.

4.1. Experimental Setup

To study the performance of the proposed approach, we have conducted various experiments. All the experiments have been carried out on a workstation with Intel Xeon QuadCore Processor, 32 GB RAM, and Ubuntu 16.04 LTS and used Keras for implementation of algorithms. We have evaluated our method by performing various experiments on benchmark data set Open Directory Project (ODP).

To perform the proposed URL classification, a total of 46280 positive samples (URLs) from Kids category have been considered. An equal number of negative samples were taken from all the remaining non-kids categories and binary classification was performed. In this research, 80% of URLs were considered for training and remaining 20% of URLs were kept aside for testing. We have performed 5-fold cross validation. We have fine tuned the parameters by varying the batch size, learning rate, optimizer, drop-out probability, activation function and number of epochs. By grid search, we fixed the optimum parameters that are listed below: a learning rate of 0.001, batch size of 128, optimizer as RmsProp with momentum 0.7, dropout rate of 0.2 and the activation function as ReLu with 50 epochs. We have fixed the maximum length as 30, the vocabulary size as 20,000 and we used GloVe for obtaining the word representation and used 300 dimensional vector. In this study, we restricted the windows size to be 3 and the number of kernels was set to 100. The objective of this proposed approach is to find out the web page whether it is Kids related or not, so we have used the accuracy as the performance metric to measure the effectiveness of the approach. The formula for calculating the accuracy is given below in Equation 1

$$Accuracy = (True\ Positive + True\ Negative) / (True\ Positive + False\ Positive + True\ Negative + False\ Negative) \quad (1)$$

where true positive denotes the number of URLs that belong to Kids category and correctly classified, whereas false positive denotes the number of URLs that are originally non-Kids, but predicted as Kids by the classifier. Similary, true negative indicates the number of URLs that are non-kids and predicted as non-kids. The false negative indicates the number of URLs that are kids, but wrongly predicted as non-kids.

4.2. Results and Discussion

As discussed in Section 3, we have designed various models to study the problem of determining the web page as kids-relevant page or not by using the URLs alone. We have carried out the experiements with the following objectives: 1) To learn the rich information present in URL and to extract the features automatically 2) To understand the patterns and the dependence between URL tokens by considering the sequence.

4.2.1. CNN based Model

The first experiment was conducted to learn the features automatically from the given data using Convolutional Neural Network. The URL was preprocessed and the tokens were extracted. As discussed in Section 1, the derived URL tokens may not be meaningful and may be misspelled. So, an automatic spelling correction was added using Symspell library. It is a symmetric Delete Spelling Correction algorithm, using which the fast spelling correction was performed for all the tokens. With this, the misspelled tokens can be corrected automatically to its nearby terms.By using CNN based approach, the model was built by training for 50 epochs with the selected hyper-parameters. For this model with the autocorrected input, an average accuracy of 0.7521 was obtained. We can conclude that, the rich features that are identified automatically by CNN along with the autospellcheck method is helpful to improve the performance.

4.2.2. RNN based Model

The dependence between the adjacent tokens in the URLs is not taken into account in the above CNN based approach. But, for identifying the category of a website, rather than considering the individual tokens separately, the dependency between the adjacent tokens could also be considered. So, it will be helpful if the information present in every token is used while processing the subsequent tokens in an URL. To achieve this, Recurrent Neural Network based experiment was carried out. However, as the simple RNN is not suitable for handling long range dependency, we have conducted the experiement with Gated Recurrent Unit. We have used the pre-trained embedding GLOVE [6] that contains 6B tokens, 400K vocab (uncased), and 300d vectors for each word. We used Glove to get the word vector for the tokens in every URL after pre-processing as mentioned in Section 3.2. The parameters used for this experiment are as follows: learning rate of 0.001, maximum sequence length as 30, drop out of 0.2, ReLu as the activation function with an RMS optimizer and binary cross entropy loss. We have obtained an accuracy of 0.7792 for the above two experiment.

In the GRU based model, we have considered the tokens in the URL only in one direction, based on the assumption that the previous tokens will also contribute for URL classification. But, the subsequent tokens may also be relevant. In the URL, the context not only depends on the past, but also on the future tokens. The combination of the previous tokens and the subsequent tokens may determine the context of the current token. So, we have used Bidirectional GRU for capturing the context in both the directions and conducted this experiment. We have obtained an accuracy of 0.7962 for this experiment.

4.2.3. Bidirectional Gated Recurrent Unit with CNN

From the above RNN model based experiments, (Section 4.2.2), we ascertained that considering the sequence information in both directions helps in improving the performance. Also, from the CNN model experiments (Section 4.2.2), it was ascertained that, rich URL features could be learnt automatically that helps in increasing the classifier performance. Hence for URL classification, rich features could be combined with sequence information and we can capture the context in both the directions. So, this experiment was conducted by combining both the models CNN and BGRU. The context rich features generated by CNN are given as input to BGRU for further classification with the same parameters listed above. For this experiment, we have obtained an accuracy of 0.8204, which is a significant improvement over all the previous models.

The summary of all the experiments are highlighted in Table 1. It could be observed that the proposed BGRU with CNN based model yielded better accuracy when compared to the other methods.

Table 1. Performance of the proposed approach

Methods	Accuracy
Baseline Experiments:	
1. CNN	0.7521
2. RNN with GRU	0.7792
3. RNN with Bidirectional GRU	0.7962
Proposed Approach : Bidirectional GRU with CNN	0.8204

5. Comparitive Study

Many researchers tried to find the solution for the problem of identifying the child related web sites by applying various techniques. Content based methods were applied by Deepshika [8] and Carsten Echoff et ål [9] and they have reported an accuracy of 63% and 72%. But, these methods will not be suitable if the contents of the webpage changes dynamically. Instead of considering the entire contents, on-page features were considered in [3]. But all the above methods are time consuming and results in waste of bandwidth. Also, it may not work if the objective is to classify the URLs on the fly.

For performing compartive analysis, we have considered the research works [5, 10] that emphasize URL based topic classification of web pages. In these existing approaches, traditional machine learning techniques such as SVM, Maximum Entropy Classifier and Naive Bayes algorithms have been applied. In [5], all - gram ($n = 4$ *to* 8) features were considered. The performance of the above method solely relies on the chosen features on the selected algorithm and has the problem in handling large scale data. In [10], a complex language based modelling approach was suggested, but there is no much improvement in the performance. Hence a deep learning based approach for URL classification was examined in this research work. By this method, there is a significant improvement in the performance. Also it has the advantage of identifying the relevant features automatically and can handle large scale data. The comparitive study is presented in Table 2.

Table 2. Comparitive Study with the Existing Approaches

Methods	Accuracy
Statistics based method [5]	0.69
All-gram features [5]	0.80
Language Model [10]	0.8109
Proposed Approach	0.8204

6. Conclusion

The usage of internet by children has increased dramatically in this digital world. The content based web page classification systems are not suitable to meet the demand of exponential growth of internet. To address this issue, URL based web page classification methodology is proposed to classify a web page as Kids-specific or not. The advantage of this approach is that, it avoids the need for downloading the entire web page contents there by saving bandwidth and eliminates the need for processing the huge amount of data. But, as the URL of a web page is a short text, extracting the suitable URL features is a challenging task. With the proposed Recurrent Connvolutional Neural Network model, the rich context aware URL token features were derived automatically by applying Convolutional Neural Network. URL can be viewed as a sequence of tokens and the neighbouring information about each token can help in classification. So this sequence information is also captured in both the directions by combinig it with a Birdirectional Gated Recurrent Unit. With the widely used ODP dataset, which is a benchmark collection for web page classification tasks,

various experiments were carried out. The performance of CNN based model and Bidirectional Gated Recurrent Unit model has been studied. From the experimental results, it is shown that, the proposed combination with a Recurrent Convloutional Neural Network performs significantly better than the existing approaches and an accuracy of 82.04% was achieved for this approach. This work can be improved by considering the character level embedding instead of using token features with word level embedding.

Acknowledgements

The authors would like to thank the management of Vellore Institute of Technology (VIT), Chennai for providing the support to carry out this research. We would also like to thank the Department of Science and Engineering Research Board (SERB), Government of India for their financial grant (Award No: ECR/2016/00484) for this research work.

References

[1] Zhang J, Qin J, Yan Q. , (2006), The role of URLs in objectionable web content categorization. In: Proceedings of the IEEE/WIC/ACM International Conference onWeb Intelligence (WI 2006); Washington, DC.
[2] Tatiana Gossen, Andreas Nurnberger (2010), Specifics of Information Retrieval for Young Users: A Survey, Information Processing and Management, 49(4):739756.
[3] Gyllstrom, Karl and Moens, Marie-Francine(2010) Wisdom of the Ages: Toward Delivering the Children's Web with the Link-based Age rank Algorithm, Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM '10, pp:159–168.
[4] Rajalakshmi, R., Aravindan, C.,(2018), An effective and discriminative feature learning for URL based web page classification. In proceedings of International IEEE Conference on Systems, Man and Cybernetics SMC 2018, Miyazaki, Japan, pp: 1374 - 1379.
[5] Baykan E, Henzinger M, Ludmila M, Weber I. A comprehensive study of features and algorithms for URL-based topic classification. ACM Trans Web. 2011;5(3):1-29.
[6] R. JeffreyPennington and C. Manning. Glove: Global vectors for word representation. 2014.
[7] Yin, Wenpeng & Kann, Katharina & Yu, Mo & Schtze, Hinrich. (2017). Comparative Study of CNN and RNN for Natural Language Processing, arXiv preprint arXiv:1702.01923, CORR.
[8] Deepshikha Patel and Prashant Kumar Singh, (2016), Kids Safe Search Classification Model, In: Proceedings of IEEE International Conference on Communication and Electronics Systems (ICCES), pp:1-7.
[9] Carsten Eickhoff, P. Serdyukov, and A.P. De Vries, (2010), Web Page Classification on Child Suitability. In Proceedings of CIKM 2010, pp:25-30.
[10] Tarek Amr Abdallah, Beatriz De La Iglasia, URL based web page classification: With n-gram language models, IC3K2014, pp: 19-33, Springer.
[11] Kan M-Y. (2004) Web page classification without the web page. In: Proceedings of the 13th internationalWorld Wide Web conference on Alternate track papers & posters (WWWAlt. '04), ACM; 2004; New York, NY.
[12] Rajalakshmi, R., Aravindan, C.: Naive Bayes approach for website classification. In: Das, V. V., Thomas, G., Lumban Gaol, F. (eds.) AIM 2011. CCIS, vol. 147, pp. 323326. Springer, Heidelberg (2011).
[13] Rajalakshmi, R., Xavier, S.: Experimental study of feature weighting techniques for URL based web page classification. Procedia Comput. Sci. 115, 218225 (2017)
[14] Rajalakshmi, R.: Identifying health domain URLs using SVM. In: Third International Symposium on Women in Computing and Informatics (WCI2015), pp. 203208. ACM (2015). https://doi.org/10.1145/2791405.2791441
[15] Rajalakshmi, R., Aravindan, C. (2018): Naive Bayes Approach for URL Classification with Supervised Feature Selection and Rejection Framework, Computational Intelligence, Wiley, pp: 363 - 396. https://doi.org/10.1111/coin.12158
[16] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A.J., & Hovy, E.H. (2016). Hierarchical Attention Networks for Document Classification. Proceedings of NAACL-HLT 2016, San Diego, California, pp: 14801489.
[17] R. Rajalakshmi, S. Ramraj and R. Rameshkannan (2019), Transfer Learning Approach for Identification of Malicious Domain Names, Security in Computing and Communications, Vol - 969, Springer,
[18] Zhou, Y., Xu, J., Cao, J., Xu, B., Li, C., & Xu, B. (2018). Hybrid Attention Networks for Chinese Short Text Classication. Computacin Y Sistemas, 21(4). https://doi.org/10.13053/cys-21-4-2847