Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 154 (2019) 62–72

Procedia Computer Science

www.elsevier.com/locate/procedia

8th International Congress of Information and Communication Technology, ICICT 2019

Web Page Classification Using RNN

Ebubekir BUBER[a*], Banu DIRI[b]

[a,b]*Computer Engineering Department, Yildiz Technical University, Istanbul, Turkey*

Abstract

Web page classification is an information retrieval application that provides useful information that can be a basis for many different application domains. In this study, a deep learning-based system has been developed for the classification of web pages. The meta tag information contained in the web page is used to classify a web page. The meta tags used are title, description and keywords. RNN based deep learning architecture was used during the tests. Transfer learning is the name given to the approach to building a machine learning model with the use of pre-trained parameters to solve a problem. The effect of using transfer learning on the system has also been examined. According to the results obtained, success rate of web page classification system is approximately 85%. It is not observed that transfer learning has significant contribution to the success rates. However, the use of transfer learning has reduced the consumed system resources.

Keywords: web page classification; classification; categorization; deep learning; RNN; transfer learning.

1. Introduction

Web pages classification (or categorization) is a machine learning problem which gets more and more important in day by day. Since the beginning of the Internet in the 90's, the number of Internet users and the number of web pages serving to users have increased at such a rapid pace and continues to grow.

* Corresponding author; ebubekirbbr@gmail.com

10.1016/j.procs.2019.06.011

Web page classification is an information retrieval application that provides useful information that can be a basis for many different application domains. Categorizing web pages provides useful information for efficient internet use, spam filtering and many other application areas. Finding relevant results quickly from millions of web sites is a serious problem that must be solved for search engines. For this reason, some search engines needed to make topic-based classification on web pages so that results returned to users could be returned better. In addition to this, web pages need to be categorized so that internet usage policies can be determined for institutions or individual uses. Web page classification can also be used by cyber security applications by blocking web pages with malicious content before they are displayed by the user. The automatic classification of web pages requires that very large amounts of data be processed. Manuel web page classification is a costly process. For this reason, manually classified web pages constitute a very small portion of the current web pages. The Web Site is the name given to the structure of a group of web pages linked to each other in various ways. A large number of web pages can be found under a website. Web pages are linked to each other.

Web page classification is the process of assigning a web page to one of the predetermined categories. Classification problem is a machine learning problem which can be solved by supervised learning approaches. Supervised learning problems consist of 2 basic stages. These are training phase and testing phase. A model can be created to solve a classification problem with a certain number of labeled samples. This model allows us to categorize web pages where we do not know about the category information.

There are also studies in which web pages are categorized by supervised methods as well as studies that are classified by unsupervised methods [1].

In unsupervised approaches, web pages are divided into groups according to similarities. Distance calculations are made to determine similarities. [1] has made an optimization of search queries by clustering search engine results.

In order to determine the category of a web page, it may sufficient to analyze only the information on the web page, however in order to be able to categorize a website, it may be necessary to analyze all web pages within that website.

Based on the assumption that the homepage of a website provides a brief summary of that website, some of the studies have only categorized the web site by analyzing the main page of the website. In this way web sites can be categorized without having to analyze many web pages.

In this study, a system was developed for the classification of web pages by using deep learning approaches. Deep learning approaches are structures with multi-layered architectures based on artificial neural networks. Deep learning approaches in many different machine learning problems in recent years have produced very successful results. Some specific deep learning approaches give good results for some types of problems. For example, Convolutional Neural Network architectures are often used for image processing. Likewise, in the case of natural language processing problems, Recurrent Neural Network based approaches can give good results. Natural language processing applications developed using CNN approaches are also found in the literature. Similarly, approaches that give good results in a specific area can be used in other types of problems. In this study, a classification mechanism was developed based on the textual properties of web pages. The system was developed using RNN architectures.

2. Problem Definition

The web page categorization is divided into several sub-titles according to the analysis outputs. Some of these titles are;

The main topic classification: Classify the web page according to the subject. For example; art, business, sports.

Functional classification: Classification of the web page according to its functional mechanism. For example; home page, login page, admin page.

Emotion classification: Classification made to understand the author's opinion in a certain way.

The application developed within this study classifying web pages doing according to main topics.

Text Classification and Web page classification problems have similarities, but they are structurally different from each other.

Firstly, the documents classified in the classical text classification problems are highly structured data sources that are well organized, semantically and structurally related between sentences and paragraphs. In these types of documents, the design of the text is completely controlled by the author. For this reason, some special characteristics can be found in writing. In this way applications such as determining the author of the text can be developed. On web pages, the structure is different from other textual data sources. It can consist of web pages, HTML contents, tags, images, audio files and videos. Additionally, textual content may not be in a semantic or structural relationship with each other. Articles about completely different topics may be found on the same web page.

Textual information such as section headings on the web page may not be a complete sentence. The web pages may contain predominantly textual content, as well as may content entirely of visual elements. Compared to the classification of classical textual documents, limited textual knowledge and a wider variety of data types may need to be analyzed in the web page classification problem.

Secondly, web pages have HTML content consisting of tag components. These HTML contents are visually presented to the users. Unlike classical textual documents, a web page has a raw output (source code) as well as a visual output (browser screen).

Finally, there are links established between web pages and other web pages or documents. It is very important to analyze hypertexts when classifying a web page. Links to other sources from the web page provide some information about the analyzed web page. This information can be helpful in classifying the web page. These are the differences between text classification and web page classification.

2.1. Literature review

The rapid development of computer and network technologies has allowed the popularity of the web to grow in a very short period of time. For this reason, the classification of web pages has become an increasingly important issue today.

There are many studies [1-3] developed for this purpose in the literature. Automatic Web Page Classification is a problem that can be solved with Supervised Learning approaches. There are many studies developed in the literature using supervised algorithms for web page classification [4-7].

Some of the traditional machine learning algorithms commonly used in the literature are as follows; Decision Trees, Artificial Neural Network [7], Support Vector Machines [6, 9], K Nearest Neighbor [10], Bayesian Algorithm [11].

In addition to these algorithms, studies on classification of web pages based on Genetic Algorithm [12, 13], Ant Colony algorithm [14] are also used in literature.

Although the web page classification problem has some structural differences from the conventional text classification problems, the algorithms and approaches used in the classification process are similar to each other. Studies on text classification in the literature have been examined, before the development of a system for the classification of web pages within this study.

As with many machine learning problems, deep learning approaches have begun to be used widely in the field of text classification in recent years. Text classification approaches developed using deep learning approaches are very important in the literature.

2.2. Deep learning approaches for text classification

There are many deep learning architectures used for text classification. Socher et al. [15] proposed Recursive Neural Network architects for sequence problems.

A sentence can be analyzed to construct a tree structure based on its semantic content in Recursive NN structure. In this structure, the performance of the model depends on the tree structure to be constructed. The creation of such a tree is a problem with O(n square) complexities (n represents the text length). For this reason, this structure is not efficient to use in long sentences and documents.

Another deep learning architecture that allows analysis of texts according to their semantic content is Recurrent Neural Network (RNN) architectures [16]. In this structure a text is processed word by word. When a word is analyzed, information about the words used before that word is also processed in hidden layers. The time complexity of this architecture is O(n). For this reason, RNN can work more efficiently than Recursive NN for long texts and documents. A disadvantage of this structure is that the RNN architecture tend to have bias. In this structure, the last word plays a more dominant role in the system than the other words. However, words with more weight can be found anywhere in the sentence or document.

Convolutional Neural Network (CNN) structures are a deep learning approach with an unbiased structure used for NLP problems [17]. In this structure, distinctive word groups can be extracted with max-pooling layers. It can be said that it is more successful than RNN constructions in catching semantic relations in this respect. Also, the time complexity of the CNN structure is O (n). This structure works by traversing a frame over the data with a certain number of lengths. Determining the frame length is an important problem. Smaller frames can cause some critical semantic information to disappear, while large frames make it difficult to train because a large number of parameters must be learned.

Recurrent Convolutional Neural Networks (RCNN) architectures and Bidirectional Recurrent Neural Network (BRNN) architectures have been proposed to overcome the shortcomings of RNN and CNN architects and to develop more robust systems. Lai et al. [18] have proposed RCNN architects to solve the problem of text classification. In this study, properties that play an important role in text classification are determined by using the max-pooling layer, and then these properties are processing with BRNN architecture. The time complexity of the RCNN approach that Lai et al. developed for text classification is also O (n).

Through deep learning approaches, successful text classification studies can be done. There are many works developed for this purpose. In this study, it is aimed to classify web pages by using deep learning approaches for text classification.

3. Web Page Classification

The application developed within the scope of this study is designed only on the classification of web pages prepared in English language and it is aimed to increase the number of languages that can be classified in the following stages.

There are many data sets available for the development of web pages classification. One of the data sets that can be used to classify Web pages is Roksit's web classification database [19]. Millions of websites are categorized in this database. Roksit uses network-based features as well as content-based information for classifying web pages. The properties used by Roksit are sufficient for a successful web page classifier. Roksit claims that the success rate of the web page classifier they built is approximately 99.9%.

Detailed information on the dataset gathered from Roksit is given in subsection 3.1. A deep learning based RNN approach has been used to web pages classification in this study. Information on the RNN architecture used in the web page classification is given in subsection 3.2. Information about word embedding is given in subsection 3.3.

Transfer learning was used, in some of the tests performed. Transfer learning allows the use of some pre-learned parameters in the deep learning architecture. The parameters used are obtained as a result of long training periods on very large data sets carried out by the researchers or technology companies. The use of these pre-learned parameters allows the training phase to be completed more quickly. Detailed information on transfer of learning is explained in subsection 3.4.

3.1. Dataset

Textual information on the target page is used in web page classification. The datasets used to classify web pages should be considered in two stages. These are;

First Stage: Only web pages and categories are included.

Second Stage: The crawled data used to classify web pages.

In the first stage, dataset has only domains and their category information. In this study, the system was developed to classify 23 categories of English web pages. These categories are given in Table 1.

Table 1. Category List

Business	Food	Game	Alcohol
Porn	Sport	News	Dating
Education	Reales	Government	Abort
Travel	Animal	Gambling	Hate
Techno	Reference	Adult	Drug
Health	Finance	Politic	

887,195 domains were used for development of the system. The category distribution of the obtained data is given in Fig. 1.

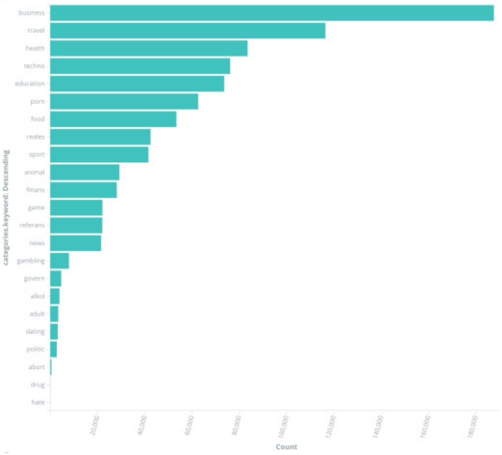

Fig. 1. Distribution of Data According to Categories

Determine what kind of data will use is necessary to build a web page classification system. A crawler module must be written to collect the information specified to classify for a web page. The information selected for use in the classification process may be information obtained from the page or information obtained from the neighboring pages [20, 21]. In this study, the information obtained from pages is processed for the classification of web pages.

The web page must have well-structured content so that a web page can successfully reach the target audience. It is very important for web pages to be indexed by search engines and be visible in the top row from the search queries to the target users.

Meta tags are one of the data types used to index the content of a web page. A well-structured web page contains some meta tags. These tags contain some information about the web page. Meta tags are entered by the web page administrator indicating the purpose / function of the web page. The most well-known and widely used meta tags are; Title, Description, Keywords.

Meta tags are included in the source code of the web page, may not be displayed on the screen by the browser to be viewed by the user.

3.2. Recurrent neural networks (RNN)

In the RNN architecture, while a word at the step of t is processed, the information of the words before the step of t is also taken as input. The basic RNN architecture consists of cells that are repeated one after the other.

A cell is taken previous cell information and given word as inputs. The recursive representation is plotted over a single cell in some references. Some other references represent the architecture as the sequential cells. The structure of the RNN architects is given in Fig. 2.

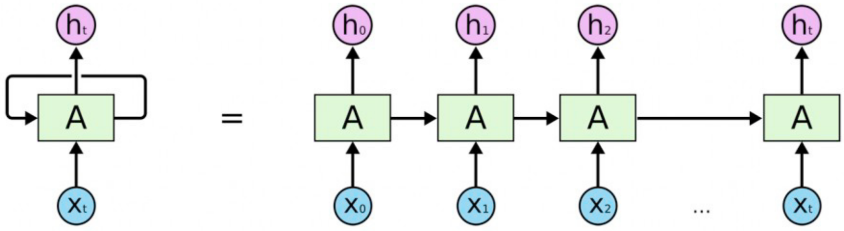

Fig 2. RNN Architecture [22]

The amount of text contained in each data instance is not specified as certain value in natural language processing problems. The sequence dimensions are reduced to a certain value so that all text can be processed.

The sequence is filled until to the specified value if the sequence size has less than the specified value. If the sequence size is more than the specified value, the excess is discarded.

3.3. Word embeddings

Word embeddings are types of word representation that has meaning for the machine learning algorithms. Word embeddings have ability that can capture semantic relationships between words or word groups. Learning of word embedding is an unsupervised learning problem. A large amount of data must be processed in order to learn successful word embeddings. The features of the given words are calculated by the algorithm performed on the large texts. The features learned represent the semantic relationships. These features are not meaningful to humans alone, although they can be understood by the algorithm.

Natural language processing problems are processed in a certain word space. Working word space is called dictionary. The words in the dictionary are in the form of a list. A special token (Unknown- UNK) is added to this list to represent words that are not in the dictionary. Only words in the dictionary can be analyzed in the algorithm for problem solving. Word vectors are used for each word in the all operations to be performed in the algorithm. The structure in which all the word vectors contained in the dictionary is called the Embedding Matrix.

Each word vector is in a column. The Embedding Matrix dimension (d x m) is for a sample with a word vector dimension d and a dictionary word count m.

The best-known methods for learning word embeddings are Word2Vec [23] and GloVe [24]. GloVe method has some performance improvements over Word2Vec method.

3.4. Transfer learning

Some machine learning applications require too many system resources to train a model. The required processing power is also increasing, if the problem complexity is high. For this reason, researchers who work on problems requiring high processing power benefit from transfer learning. Thus, they can build successful deep learning architects with less effort.

Transfer learning is based on the use of pre-learned parameters which are calculated by someone else. Transfer learning allows many researchers to save time and system resources in many fields of application. In situations where there is not enough data to be used during the training phase, learning transfer has a positive effect on the success of the system.

Transfer learning is used in many fields of application as well as in the field of natural language processing. Learning word embeddings is a costly process. It is necessary to do learning on very large texts for learning word embeddings. In order to learn word embeddings successfully, collecting large amounts of data is a very difficult problem in itself, processing this data is also a difficult and time-consuming problem. Some leading technology firms, some universities and researchers in the field are able to share the parameters they have obtained by using these processes, which require a lot of time and effort, to be used by other researchers. In this regard, researchers who do not have sufficient hardware resources can work on well-learned models.

4. Experimental Results

In this study, the information obtained from the web pages was used to classify the web pages. There are also studies where information about neighboring web pages is used to classify a web page [20, 21].

Meta tags contain information about the purpose and content of a web page. In this study, meta tags were used to classify web pages. The meta tags used are title, description and keywords. RNN architecture was used during the tests.

The tests were performed on a leased GPU server from Floydhub [25]. The hardware features available on the server are; Tesla K80 GPU (1 processing Unit) 12 GB VMemory, 61 GB RAM, 4 core Intel(R) Xeon(R) CPU E5-2686 v4 @ 2.30GHz processor, 100 GB SSD Disk.

During the tests, two cases were analyzed. One of the tests is performed using transfer learning, the other one is performed without using transfer learning. The results obtained from the tests were evaluated. Tests were carried out using Keras framework [26]. Keras default values of hyperparameters which are not mentioned below is used for the tests. The hyperparameters used for the tests are;

Epoch count: 5

Loss function: Categorical Cross Entropy

Maximum sequence size: 100

Parameter update function: ADAM; Learning rate: 0.01, Beta_1: 0.9, Beta_2: 0.999.
Overall sample count: 887,195
Validation set split ratio: 0.2
Unique word count in dictionary: 1,387,699
All layers were created with 128 neurons.
Word embedding dimension: 100
Batch size: 1000

5-layered RNN architecture was used with LSTM cells during the tests. Learning word embeddings is a process that requires high processing power. In this experiment, transfer learning was performed using Stanford University's word vectors [27], which were calculated using the Glove method. The word embeddings used in transfer learning were created by analyzing 6 billion words in corpus of Wikipedia 2014 + Gigaword 5. There are 400,000 unique words in embedding matrix used. Same test was also examined without using transfer learning, then the results obtained were evaluated.

While the total number of unique words in our corpus is 1,387,699 only 171,029 words can be processed with deep learning architecture. The other words are not included in the embedding matrix which has 400,000 unique words in it. In this case, very few of the words in our dictionary can be processed. The success rates of the test performed using transfer learning and without using transfer learning are given in Fig 3.

Although the 171,029 words can be processed, the success rate the success rate obtained is up to 86% in the test using transfer learning. In the test, which does not use transfer learning, word embeddings are learned for all 1,387,699 words in the dictionary. Achieved success rate is approximately 85%. (TL: Transfer Learning was used. - Non-TL: Transfer Learning was not used.)

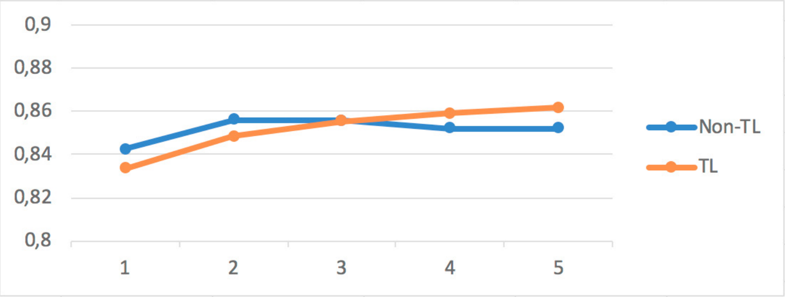

Fig. 3. Validation Accuracies for the Comparison of TL and Non-TL

Values for the average running times for five epochs of the tests are given in Fig 4. The test using learning transfer was completed faster than the other test. The given values are in minutes format. Success rates obtained in the last epoch for TL and Non-TL tests are respectively as follows 86,18 %, 85,21%.

Learning word embeddings is consumed a lot of system resources. The consumption of GPU Resources for Non-TL test if given in Fig. 5 and the consumption of GPU Resources for TL test is given in Fig. 6. The graph given in Fig.5 and Fig. 6 is updated every minute.

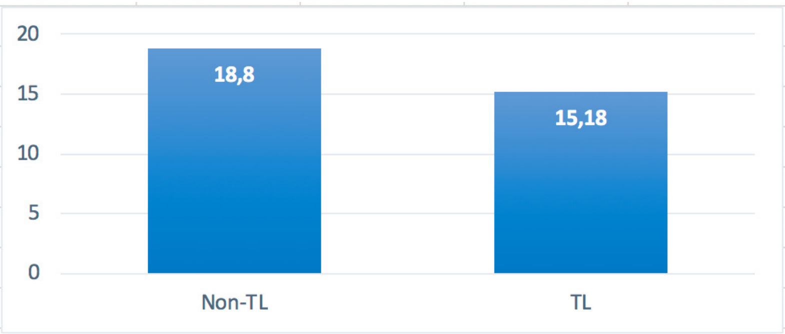

Fig. 4. Running Times for the Comparison of TK and Non-TL

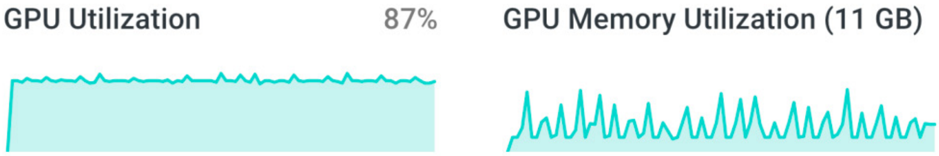

Fig. 5. The Consumption of GPU Resources for Non-TL Test

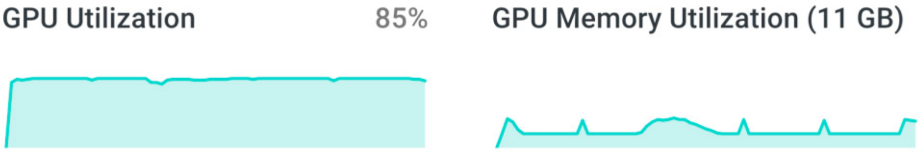

Fig. 6. The Consumption of GPU Resources for TL Test

In the Non-TL test, the processing power consumption is somewhat higher than the other test. There is a big difference for memory consumption between TL test and Non-TL test. memory consumption varies in a much wider range in Non-TL test. Table 2 shows some statistical values related to memory consumption. The values given represent consumption percentages of 12 GB GPU Memory.

Table 2. Statistical Values for Memory Consumption

Tests	Min	Max	Average	Standard Deviation
TL	18.0%	38.0%	21.64%	17.13%
Non-TL	16.0%	77.0%	31.58%	6.73%

According to the results obtained, success rates in TL and Non-TL tests are close to each other. However, the Non-TL test runs slower than the other test and consumes more GPU resources. Transfer learning has not affect a significant contribution to the achievement values. It is thought that because there is a large number of data samples for web page classification in this study.

5. Conclusion and Future Work

Web page classification is an information retrieval application that provides useful information that can be a basis for many different application domains.

Some search engines needed to make topic-based classification on web pages so that results returned to users could be returned better. In addition to this, web pages need to be categorized so that internet usage policies can be determined for institutions or individual uses. Web page classification can also be used by cyber security applications by blocking web pages with malicious content before they are displayed by the user.

A deep learning-based system has been developed for the classification of web pages in this study. The data obtained from Roksit was used in the developed system. The meta tag information contained in the web page is used to classify a web page. The meta tags used are title, description and keywords. This information has been collected by the crawler module developed in this study. RNN based deep learning architecture was used during the tests. The effect of using transfer learning on the system has also been examined. The tests were performed on a leased GPU.

According to the results obtained, success rate of web page classification system is approximately 85%. It is not observed that transfer learning has significant contribution to the success rates. However, the use of transfer learning has reduced the consumed system resources. It is thought that the reason of that there are large number of samples used in the development of the system. The tests for verification of this interpretation will be performed in further studies. In future studies, hyperparameter optimizations will be also performed to increase success rates.

6. Acknowledgements

Thanks to Roksit for contributing to the development of this study.

7. References

1. Dural Burak, Türkçe Arama Motoru Sonucu Kümeleme Çalışmaları, Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği, Yayımlanmamış Yüksek Lisans Tezi, İstanbul.
2. Altıngövde, I_. S., Özel, S. A., Ulusoy, Ö., Özsoyoglu, G., & Özsoyoglu, Z. M. (2001). Topic-centric querying of Web information resources. Lecture Notes in Computer Science, 2113, 699–711.
3. De Bra, P. M. E., & Post, R. D. J. (1994). Information retrieval in the World Wide Web: Making client-based searching feasible. Computer Networks and ISDN Systems, 27(2), 183–192.
4. Menczer, F., Pant, G., & Srinivasan, P. (2004). Topical Web crawlers: Evaluating adaptive algorithms. ACM Transactions on Internet Technology, 4(4), 378–419.
5. Qi, X., & Davison, B. D. (2009). Web page classification: Features and algorithms. ACM Computing Surveys, 41(2) (article 12).
6. Chen, R. C., & Hsieh, C. H. (2006). Web page classification based on a support vector machine using a weighted vote schema. Expert Systems with Applications, 31, 427–435.
7. Selamat, A., & Omatu, S. (2004). Web page feature selection and classification using neural networks. Information Sciences, 158, 69–88.
8. Wakaki, T., Itakura, H., Tamura, M., Motoda, H., & Washio, T. (2006). A study on rough set-aided feature selection for automatic web-page classification. Web Intelligence and Agent Systems: An International Journal, 4, 431–441.
9. Rung-Ching Chen, Chung-Hsun Hsieh (2006), "Web Page Classification based on a support Vector Machine using a weighted vote schema", Expert Systems with Applications, Vol. 31, Issue 2, pp:427-435.
10. Ozel, S. A. (2011, June). A genetic algorithm based optimal feature selection for web page classification. In Innovations in Intelligent Systems and Applications (INISTA), 2011 International Symposium on (pp. 282-286). IEEE.

11. Sara Meshkizadeh, Amir Masoud Rahmani, Mashallah Abassi Dezfuli (2010), “Web Page Classification based on URL features and Features of Sibling Pages”, IJCSIS, Vol. 8, No:2.
12. Ozel, S. A. (2011). A web page classification system based on a genetic algorithm using tagged-terms as features. Expert Systems with Applications, 38(4), 3407-3415.
13. Ribeiro, A., Fresno, V., Garcia-Alegre, M. C., & Guinea, D. (2003). Web page classification: A soft computing approach. Lecture Notes in Artificial Intelligence, 2663, 103–112.
14. Nicholas Holden and Alex A. Freitas, (2004), “Web Page Classification with an Ant Colony Algorithm”, Parallel Problem Solving from Nature, LNCS, Springer, Vol.3242, pp:1092-1102.
15. Richard Socher, Q Le, C Manning, and A Ng. Grounded compositional semantics for finding and describing images with sentences. In NIPS Deep Learning Workshop, 2013.
16. Medsker, L. R., & Jain, L. C. (2001). Recurrent neural networks. Design and Applications, 5.
17. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; and Kuksa, P. 2011. Natural language processing (almost) from scratch. JMLR 12:2493–2537.
18. Lai, S., Xu, L., Liu, K., & Zhao, J. (2015, January). Recurrent Convolutional Neural Networks for Text Classification. In AAAI (Vol. 333, pp. 2267-2273).
19. Domain Categorization, https://www.roksit.com/domain-categorization/, accessed on October 2017
20. Wongkot Sriurai, Phayung Meesad, Choochart Haruechaiyasak (2010), “Hierarchial web page Classification based on a Topic Model and Neighboring Pages Integration”, International Jopurnal of Computer Science and Information Security, Vol. 7, No.2.
21. Sara Meshkizadeh, Amir Masoud Rahmani, Mashallah Abassi Dezfuli (2010), “Web Page Classification based on URL features and Features of Sibling Pages”, IJCSIS, Vol. 8, No:2.
22. Understanding LSTM Networks, http://colah.github.io/posts/2015-08-Understanding-LSTMs/, Son Erişim: Mayıs 2018.
23. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
24. Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).
25. FloydHub is a zero setup Deep Learning platform for productive data science teams. https://www.floydhub.com/, Son Erişim: Nisan 2018.
26. Keras: The Python Deep Learning library, https://keras.io/, Son Erişim: Mayıs 2018.
27. GloVe: Global Vectors for Word Representation, https://nlp.stanford.edu/projects/glove/, accessed onApril 2018.