An Evolution and Portrays of Web Page Classification

Abstract- Net created explosive information to grow for classification purposes. It contains websites, net crawl, promotion, data extraction, built-in, recovery, management instructions, filter, based unwanted data, hazardous database content, parental management-based system. As the memory allocation area has been growing, the latest computer system performance has been developed alongside the expert on machine learning models in the texting and imaging classification. For several researchers, online disadvantage about page classification begins and focuses. The web classification remains automatically due to the quality and diversity of the net pages at an early stage. The image, text and hyperlink sizes and the machine price are varied. In this survey, we only projected methology within the literature survey, but jointly traced development and showed completely different views on this inconvenience.

We have studied and invented I discourse data closing the term square measurement in classification matter mainly unnoticed.(ii) hyperlink to the structure, and distribute square measures in classification materials in the form of texts, tags or hyperlinks.(iii) measure, measure, and accuracy of each feature classification within distinct gap analysis, measure the effect option to identify web content categories.(iv) heavy computational trust strategies classification, problem based analytical extraction feature. (v), despite the understudy, learn to supervise large quantity untagged web content, thus enormous price markings, the importance of Web content owing to classification. (vi) deep learning, continuous N/W, strengthening learning remain under study but the Web Content Classification remains intriguing. (vii) establish an analysis of common location metrics, develop close work and classify stay benchmarks by access to the web content.

Key words- *Web page classification, Image classification, Text classification, Deep learning, Machine learning, Artificial intelligence.*

INTRODUCTION

Planet wide Internet (called Web) is unimaginably growing, which makes it difficult to search for websites that convey satisfactory information and separate undesirable or harmful contents. Web sites, such as scam, phishing, violence, radicalism, cyber-threats, porn, etc., have proliferated in the last decade. On the other hand, the enormous number of websites featuring different topics has hampered the retrieval of information and the extraction of models to provide optimal results. In connection with specialization of machine-learning models for text and image classification, explosive performance and memory housing growth in computer equipment have provided the means for automatically solving difficult linguistic problems, which seemed too unreal for any computer for more than ten years. Another problem is that the websites supported their content are linguistic classified. Website classification means that the way a website is distributed to one or more defined website classes, which play an important role in centred slopes, help to develop Internet directories, to analyse the topic-specific internet links, advertise speeches and to analyse the topical structure of the Internet. The quality and seriousness of this disadvantage combined with a variety of website classification perspectives and models led the US to conduct an examination of literature on this topic, and to take light on its challenges and gaps. The rest of this paper is arranged accordingly. Section two criticises the progressive website classification literature. This section explores website classification models one by one below three main text, image and combined method classes. Section three offers a collective picture of current models and analysis lacunae, whereas Section four focuses on major website classification limitations and offers potential gateways. Section 5 ends with a summary of analytical deficiencies and future directions.

WEB PAGE CLASSIFICATION METHODS

In this section, the methods for classing web pages in a literature are explored withinText-based grouping (a), picture-based grouping (b), and text-based and picture-based ▪grouping (c). Thus, fundamental differences in how features for machine learning are extracted from text and image motivate automatic webpage classification methods. Section 3 however provides for the classification of website classification methods on the basis of the applicable classification and accuracy. Tokenisation, non-alphanumeric deletion of characters, deletion of stop words, conversion and stemming are pre processing steps of all text-based methods of classification.

i) Text-based classification - Hosts and provides sexual services such as escorts, adult entertainment, massage services, and so on. Hosts or escorts advertisements. It is in the interest of both criminal and business courts to detect such contents. Because of their illegality, law enforcement agencies would be able to detect such content. Companies are interested in finding such content, in order to protect minors' exposure and report on their platform illegal activity. Thus, the detection of online pornography has become a major focus of the literature. To detect those web pages, 12 binary characteristics extracted from the text were used, Alvariet others used a semi-supervised SVM.

The binary functionality is:

- third-party speech,
- first-person plural pronouns,
- high entropy content,
- three 4-g (found in three binary functions),
- interesting words and phrases
- They took people from countries of interest on escorts
- included several victims,
- a poor victim,
- references to infamous websites,
- and a mention to spa massage therapy.

ii) Image-based classification- The most typical method for analysing the visual content of pornographic websites, which necessitates extensive analysis and computations. has been the extraction of features from skin regions. The skin colour pixels in a picture with colour and Texture Information are recognised by traditional pornographic image recognition techniques. Gaussian blend models (GMM), fitted to the histogram of the skin, Since skin has a distribution with several peaks in the colorspace, had been a popular detection technique for skin pixels.
The expectation-maximisation approach for obtaining mixed model parameters, on the other hand, has substantial computational expenses. In order to reduce processing expenses, Hu et al. adopted adaptive bin sizes in the histogram of skin colours. Instead of using GMM, Ahmadi et al. utilised a neural network with one hidden layer to detect whether a pixel was skin. The neural network's entrances are the colour properties of the pixel and its four neighbours.
iii) Classification based on both text and images - In this part, we'll look at some attempts to combine textual and visual data for web page classification. All of these combined techniques outperformed individual image or text classifiers in terms of precision. Fake websites make illegal money by misrepresenting themselves as legitimate sources of information, commodities, or services. There are two types of fraudulent website detectors: Survey and categorization systems are two types of systems. Lookup systems identify bogus websites by comparing their URLs to a blacklist of fake URLs compiled by system users and online communities. Lookup systems are quick to find URLs with few false positives, but they are sluggish to blacklist new phoney websites (since blacklists rely only on user reports of bogus URLs). Classifier's systems detect bogus websites based on the function of their contents, body text, page style, and other factors, photos, picture hash, scan for encryption, domain name, host land, registration date, and URLs. Classifier systems are generally less precise than search systems and slower than search systems because it takes longer to

classification of a website than to search for a blacklist URL. However, new fake websites are proactive and faster to blacklist.

OUTLINE AND RESEARCH GAPS

Table 1 provides an overview of studies on the Web page with the features and the method of reduction of dimensionality. Table 2 summarises Website classification studies, including datasets, data size, class size, classification method, assessment method, and classification accuracy. The studies in each table are arranged in descending order of categorization accuracy. Table 2 shows the categorization accuracy in F1, the most common in the literature (also known as the F- or F-score). We estimated F1 in studies where F1 was lacking but the confusion matrix was present. In the absence of both the F1 and the confusion matrix, the overall accuracy (OA) is provided in a study Figure 1. Provides a summary of the research based on the classification method employed and the precision with which the classification was made. Only the highest precision in a study is displayed in Fig. 1. If different classification accuracies were obtained using a method,

Method	T	CIT	H	SI	M	I	V	CI	SS	Dimensionality reduction method
[101]	✓					✓				
[66]						✓				
[35]	✓					✓				
[69]							✓			
[5]	✓		✓		✓				✓	
[91]						✓				
[34]	✓	✓				✓				
[61]							✓			
[83]						✓		✓		Information-gain-based feature selection for contextual features
[53]	✓									
[76]	✓									PCA
[68]	✓			✓						
[4]	✓		✓		✓	✓				
[25]	✓	✓								
[56]	✓					✓				
[11]						✓				
[51]	✓			✓						
[85]						✓				
[42]	✓								✓	
[8]	✓	✓				✓				Autoencoder for visual features
[2]	✓				✓	✓				
[54]	✓									Autoencoder
[73]						✓				

Table 1 Different classification studies of websites, their input features and their method of dimension reduction (T: classification of textual content; CIT: Consideration of text classification contextual information; H: Consideration of hyperlinks; IS: Structural information consideration, i.e. HTML tags; M: metadata consideration I: content-based grading of images; V: video-based classification; CI: context-based grading of images; SS: semi-training supervised).

Figure 2 depicts each of them. The frequency of the classification approach, alone or in combination with other classification methods, in the literature evaluated. The following broad observations can be drawn from the above table and figures. The grading of websites was primarily centred on pornographic websites until around a decade ago. Beyond the detection of pornographic websites and textual elements, the apps have evolved and become a fundamental part of Web page classification. Table 1 reveals that seven research classify web pages using images and texts; ten studies ignore the images and eight ignore the text. However, no particular models on how the classification accuracy disregards either the images or the text are seen. Website classes can be classified and the websites

indicated via the hyperlinks, the results of which can be incorporated into the classification of the original website. More sophisticated approaches to Videos, photos, text, and hyperlinks that are incompatible with a web page's theme may be created and then removed. More emphasis has recently been paid to the increasing expansion of unlabeled web data and the variety of traits that can be retrieved in recent studies. However, we were unable to locate any research that evaluated the effectiveness of features in distinguishing between Web page classes, or in other words, the contribution of each feature to classification accuracy. Filtering away linked traits or features is a critical effort, which contribute very little to none difference between classes. The best accuracy is achieved, according to table 2, in the use, alone or together with other classifiers, of kNN.

Method	Dataset	Number of samples	Number of classes	Classification method	Evaluation method	Accuracy%
[101]	Phishing and non-fishing Web pages	~9000	2	• Naïve Bayes for text classification • Nearest neighbor for image classification	One-fold	F1 = 99~100
[66]	Pornographic and non-pornographic Web pages	48,600	2	• CNN	One-fold	OA = 99
[35]	Pornographic and non-pornographic Web pages	3090	2	• Random forest for image classification • kNN-based and SVM-based methods for text classification	One-fold	F1 = 98~99
[69]	Pornographic and non-pornographic Web pages	1000	2	• CNN	Two-fold	OA = 98
[5]	Human trafficking and non-human trafficking Web pages	20,000	2	• SVM	Ten-fold	F1 = 95
[91]	Pornographic and non-pornographic Web pages	150,000	3	• CNN	One-fold	F1 = 95
[34]	Pornographic and non-pornographic Web pages	~4000	3	• Naïve Bayes for text classification • Nearest neighbor for image classification	One-fold	F1 = 94
[61]	Pornographic and non-pornographic Web pages	800	2	• CNN	Five-fold	OA = 94

[83]	Web page categories of knife, crab, people, airplane, vessel, grapes, and gun	1400	7	• SVM	Three-fold	F1 = 92
[53]	Web page categories of culture, education, entertainment, finance, health, religion, government, science, sport, and travel	17,431 18,099 19,444	10	• Naïve Bayes	One-fold	OA = 93 OA = 92 OA = 91
[76]	Yahoo sports news Web page categories	4096	12	• Neural network	One-fold	F1 = 90
[68]	Course-related and not related Web pages	1051	2	• Genetic algorithm	One-fold	F1 = 91
	Conference and non-conference Web pages	292				F1 = 90
	Student and non-student Web pages	5405				F1 = 69
[4]	Pornographic and non-pornographic Web pages	1585	3	• Decision tree for text classification • Neural network for image classification	One-fold	F1 = 87
[25]	Yahoo Web page categories	10,000	5	• Multinomial Naive Bayes	Ten-fold	F1 = 89
				• SVM		F1 = 89
				• 1 nearest neighbor		F1 = 87
				• 5 nearest neighbors		F1 = 87
				• Naive Bayes		F1 = 85
				• Decision Tree		F1 = 81
				• Bayesian Network		F1 = 81
[56]	News article Web page categories	1043	4	• Random forest for text classification • Random forest for image classification	One-fold	F1 = 85
[11]	Pornographic and non-pornographic Web pages	11,005	2	• Neural network	Ten-fold	F1 = 83
				• k-nearest neighbors		F1 = 82
				• SVM		F1 = 82
				• Generalized linear model		F1 = 73
[51]	Health-related and not related Web pages	~1000	2	• Simplified swarm optimization	One-fold	F1 = 81
	Computer-related and not related Web pages	~1000				F1 = 78
	Science-related and not related Web pages	~1000				F1 = 78
	Art-related and not related Web pages	~1000				F1 = 73
[85]	Child-pornographic and non-child-pornographic Web pages	2000	2	• SVM	Five-fold	OA = 79
[42]	Academic Web page categories	4199	4	• Dirichlet mixture distribution	One-fold	F1 = 77
	Newsgroup Web page categories	19,946	20			F1 = 77
[8]	Tweet polarities (Sanders Corpus dataset)	3625	2	• Logistic regression (textual and visual features are concatenated)	Ten-fold	F1 = 73
	Tweet polarities (Sentiment140 dataset)	1,700,000			One-fold	F1 = 83
	Tweet polarities (SemEval-2013 dataset)	13,434			One-fold	F1 = 73
	Tweet polarities (SentiBank Twitter dataset)	6,03			Five-fold	F1 = 57
[2]	Fake and non-fake Web pages	1400	2	• SVM (textual and visual features are concatenated)	One-fold	F1 = 71
[54]	Social comment emotions (SinaNews dataset)	1246	6	• Deep neural network	One-fold	F1 = 60
	Social comment emotions (ISEAR dataset)	7666	7			F1 = 51
	Social comment emotions (SemEval-2007 dataset)	4570	8			F1 = 39
[73]	Pornographic and non-pornographic Web pages	69,260	2	• SVM	One-fold	F1 = 19

Table 2 Different studies of web pages, together with their data set and classification method, in the following order.

Tree is the most precise classification system. In different studies, SVM is the most commonly used method that achieves a wide variety of accuracies, ranging from 19% to 95%. Figure 1 shows the picture classification model CNN, which achieved the highest level of accuracy. According to Fig. 2, the SVM classification is the most extensively used, followed by neural networks, kNN, random forests, and decision-making tree. One of the techniques to web page classification that hasn't changed is to strengthen learning.

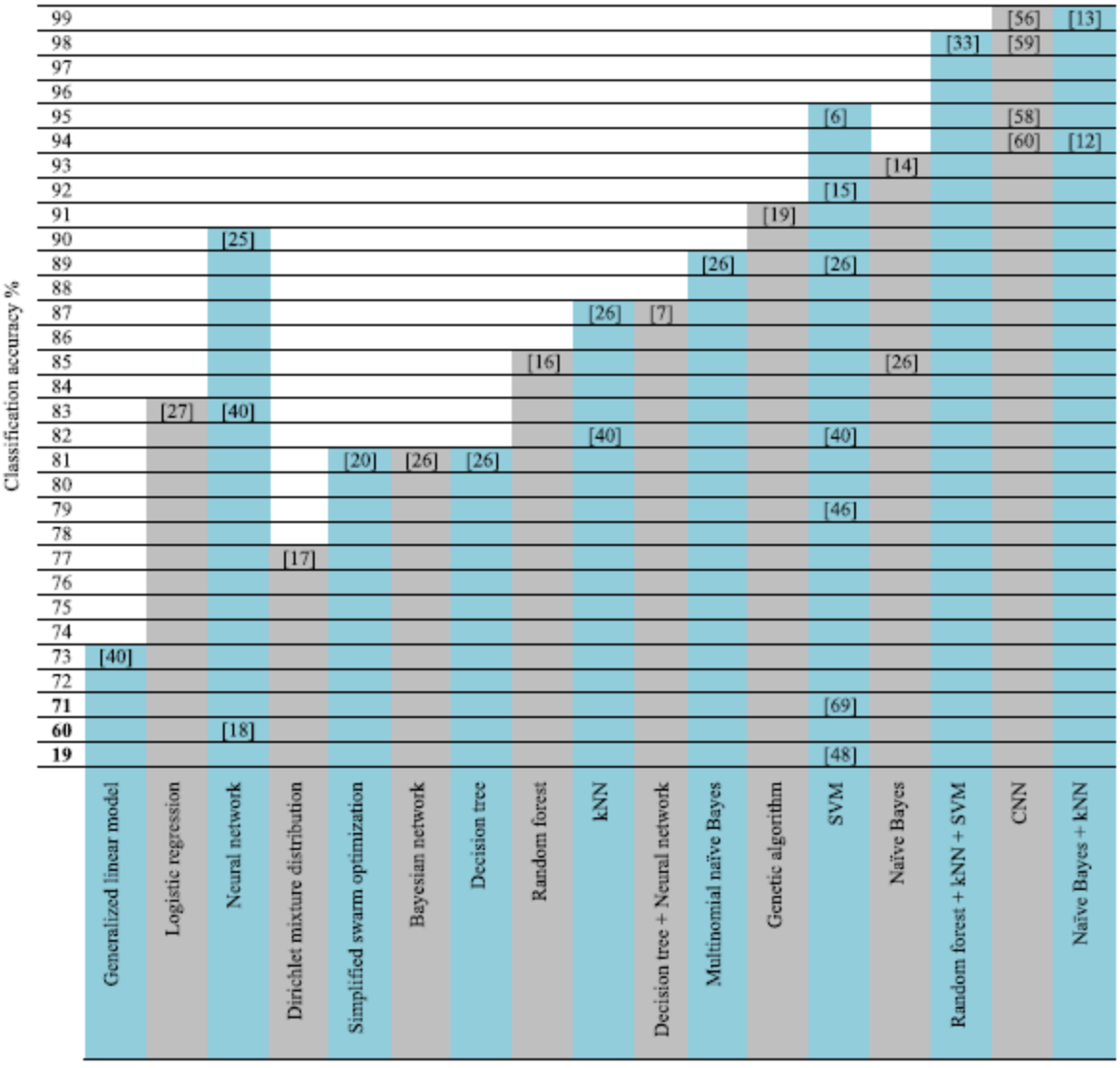

Fig. 1 Different Web page classifiers and their classification accuracy

LIMITATIONS AND POTENTIAL DOORWAYS

This section highlights some of the limitations before classifying automatic web pages. It also presents solutions used to meet similar literary challenges. Tables and figures in Section3 do not mention the studies reviewed in this section as they do not focus specifically on classification of websites.

i) Limited training data- The small number of marked web pages presents a challenge for the efficient training of automatic web page classificators. A potential solution to this problem would be to find a way to use unlabeled web pages in a semi-controlled classification process. The labelled samples would, if possible, be multiplied by representing Web pages in various feature spaces. Both

solutions were applied by Fakeri- Tabrizi et al. to overcome the lack of enough images labelled for classification training. They proposed a self-learning approach, using multi-view displays to produce pseudo-labeled training information when a large number of unlabeled samples are available.

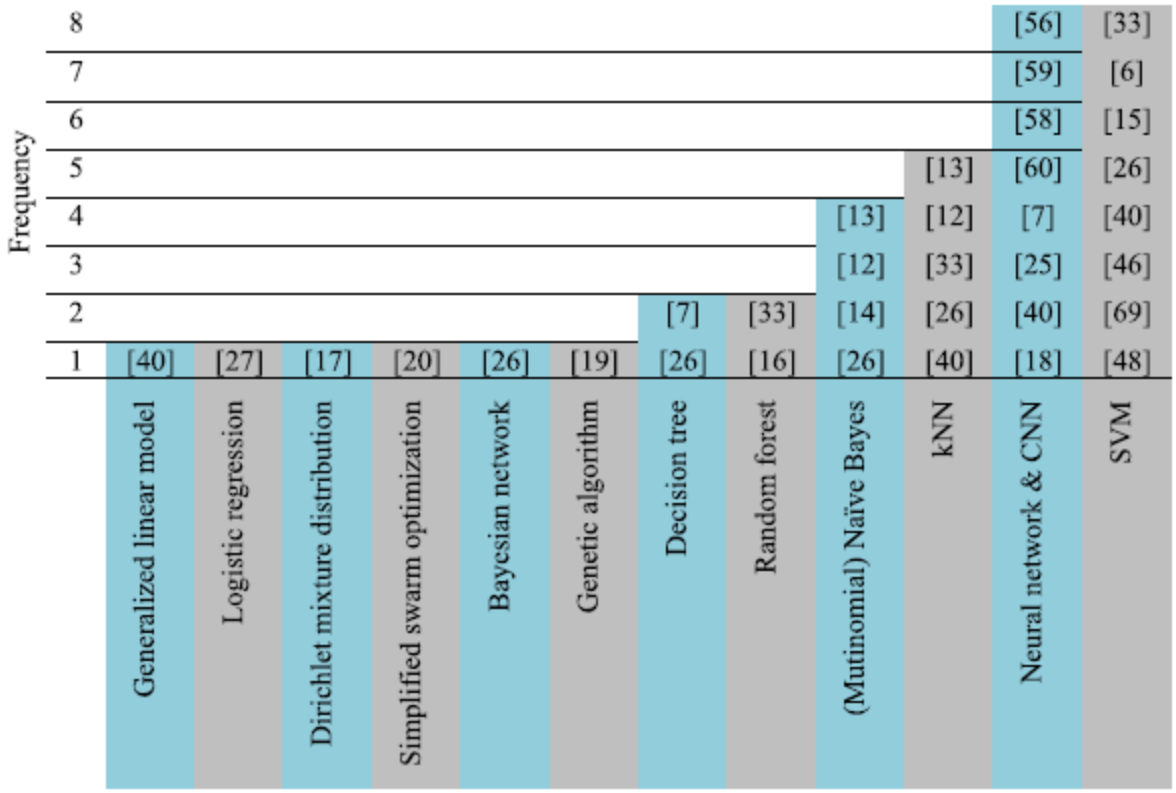

Fig. 2 each classification method's frequency in the reviewed literature, alone or in combination with other classification methods

ii) Overlooking structural information- A webpage is a tree with a structural element representing each node. A node includes two types of information: information on the tag or label and information on the content. The content of a node with the paragraph of the label, for example, is the text of the paragraph.

While the rating of the web page focuses primarily on flat content, the structural information has not been given much attention. How structural information can be effectively applied in web page classification needs to be thoroughly studied.

iii) Limitations in visual feature extraction and their classification accuracy - Low precision and complicated handcrafted characteristics are two challenges for the classification of web pages visual content. Due to its independence from handcrafted visual features and excellent abstract and semantine ability, CNN has been shown to exceed other image classification approaches in some applications. However, in the classification of webpages CNN was not appropriately applied mainly for two reasons: it needs up to millions of training pictures, and receives only the same-sized Images as input. In order to overcome the limited number of training images, pre-training and data increase via translation, horizontal reflection and intensity alterations in RGB channels have been widely proposed.

iv) Overlooking the sequence of terms in the textual content- Text is inherently sequential, as understanding the preceding words contributes to understanding the following words. In the current web page classification literature, the sequence of terms in textual content has been overlooked. As previously mentioned, term frequency vectors lose the text structure by breaking it down into terms and its frequencies. Although some researchers have proposed changes to term frequencies to take partly contextual information into account, its scope is limited and cannot easily be extended to large datasets. Recurrent neural network (RNN) is a possible solution to this challenge, offering the

possibility to classify the complete text while taking into account both the sequence of the arbitrary stream of textual data and its contextual information.

v) Lack of a detailed test bed- Most studies use self-collected data sets to train and test the classification of their website. Without a comprehensive test bed, the accuracy of various web pages classifiers is difficult to compare. However, the development and establishment of the standard benchmarking of a detailed test sheet remain a gap in the evaluation of web page classifiers.

CONCLUSIONS AND FUTURE DIRECTIONS

In summary, the following limits and possible direction are highlighted in our study:

a) Classifying the textual content:

- In the classification of text content, metadata and contextual data relating to terms are mainly neglected.
- Structure and distribution in HTML tags and hyperlinks are under-studied in the classification of textual content.

b) Classifying the visual content:

- Methods for image classification are largely based on computer-intensive and problematic feature-specific analyses, for example the detection of pornographic pictures rely on extracted features not appropriate for other image types.

c) Classifying the Web page as a whole:

- A significant research gap is that feature efficiency is measured in distinguishing between Web pages or in measuring every feature's contribution to classification accuracy.
- Despite the importance in the Web page classification of semi-monitored learning, it is understood due to the enormous number of unlabeled web pages and the high labelling cost.
- Classification of websites Deep learning, CNN, RNN and enhancement learning remain underexplosed but intriguing.
- Develop a detailed test bed together with assessment metrics and set standards

The criteria remain a gap in the evaluation of classifications of web pages.

REFERENCES

[1]. Abbasi, A., & Chen, H. (2007). Detecting fake escrow websites using rich fraud cues and kernel-based methods. 17th Annual Workshop on Information Technologies and Systems, (pp. 55–60). Montreal,Canada.

[2]. Abbasi A, Chen H (2009) A comparison of tools for detecting fake websites. Computer 42(10):78–86

[3]. Abin, A. A., Fotouhi, M., & Kasaei, S. (2008). Skin segmentation based on cellular learning automata. 6th International Conference on Advances in Mobile Computing and Multimedia (pp. 254-259). Linz, Austria:ACM.

[4]. Ahmadi A, Fotouhi M, Khaleghi M (2011) Intelligent classification of web pages using contextual andvisual features. Appl Soft Comput 11(2):1638–1647

[5]. Alvari H, Shakarian P, Snyder JK (2017) Semi-supervised learning for detecting human trafficking.Security Informatics 6(1).

[6]. Ap-Apid, R. (2005). An algorithm for nudity detection. 5th Philippine Computing Science Congress, (pp.201-205).

[7]. Arentz WA, Olstad B (2004) Classifying offensive sites based on image content. Comput Vis Image Underst 94(1–3):295–310

[8]. Baecchi C, Uricchio T, Bertini M, Bimbo AD (2016) A multimodal feature learning approach for sentiment analysis of social network multimedia. Multimed Tools Appl 75(5):2507–2525

[9]. Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural emachine translation by jointly learning to align and translate. arXiv preprint , arXiv:1409.0473.

[10]. Bengio Y, Simard P, Frasconi P (1994) Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Netw 5(2):157–166

[11]. Bosson, A., Cawley, G. C., Chan, Y., & Harvey, R. (2002). Non-retrieval: blocking pornographic images. International Conference on Image and Video Retrieval (pp. 50-60). Berlin, Heidelberg: Springer.
[12]. Chan, Y., Harvey, R., & Bangham, J. A. (2000). Using colour features to block dubious images. 10th European Signal Processing Conference. 3, pp. 1-4. IEEE.
[13]. Chiu, J. P., & Nichols, E. (2015). Named entity recognition with bidirectional LSTM-CNNs. arXiv preprint , arXiv:1511.08308.
[14]. Chou N, Ledesma R, Teraguchi Y,Mitchell JC (2004) Client-side defense against web-based identity theft. In: 11th annual network and distributed system security symposium. Internet Society, San Diego
[15]. Chua CE, Wareham J (2004) Fighting internet auction fraud: an assessment and proposal. Computer 37(10):31–37
[16]. Collobert R,Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P (2011) Natural language processing (almost) from scratch. J Mach Learn Res 12(Aug):2493–2537
[17]. Denoyer L, Gallinari P (2004) Bayesian network model for semi-structured document classification. Inf Process Manag 40(5):807–827
[18]. Diligenti, M., Gori, M., Maggini, M., & Scarselli, F. (2001). Classification of html documents by hidden tree-markov models. Sixth International Conference on Document Analysis and Recognition (pp. 849- 853). Seattle, WA, USA: IEEE.
[19]. Du, R., Safavi-Naini, R., & Susilo, W. (2003). Web filtering using text classification. The 11th IEEE International Conference on Networks (pp. 325-330). IEEE.
[20]. Dumais, S., & Chen, H. (2000). Hierarchical classification ofWeb content. 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 256-263). ACM.
[21]. Fakeri-Tabrizi A, Amini M-R, Goutte C, Usunier N (2015) Multiview self-learning. Neurocomputing 155(1):117–127
[22]. Farfade, S. S., Saberian, M. J., & Li, L.-J. (2015). Multi-view face detection using deep convolutional neural networks. 5th International Conference on Multimedia Retrieval (pp. 643-650). ACM.
[23]. Fauzi F, Belkhatir M (2010) A user study to investigate semantically relevant contextual information of WWW images. International Journal of Human-Computer Studies 68(5):270–287
[24]. Fauzi F, Belkhatir M (2013) Multifaceted conceptual image indexing on the world wide web. Inf Process Manag 49(2):420–440
[25]. Fersini E, Messina E, Archetti F (2008) Enhancing web page classification through image-block importance analysis. Inf Process Manag 44(4):1431–1447
[26]. Forsyth DA, Fleck MM (1999) Automatic detection of human nudes. Int J Comput Vis 32(1):63–77
[27]. Hammami, M., Chahir, Y., & Chen, L. (2003).WebGuard: web based adult content detection and filteringsystem. IEEE/WIC International Conference on Web Intelligence (pp. 574-578). IEEE.
[28]. Hammami M, Chahir Y, Chen L (2006) Webguard: a web filtering engine combining textual, structural, and visual content-based analysis. IEEE Trans Knowl Data Eng 18(2):272–284
[29]. Hashemi M, Hall M (2018) Visualization, feature selection, machine learning: identifying the responsible group for extreme acts of violence. IEEE Access 6(1):70164–70171
[30]. Hashemi, M., & Hall, M. (2018). Identifying the responsible group for extreme acts of violence through pattern recognition. International Conference on HCI in Business, Government, and Organizations (pp. 594-605). Cham: Springer.
[31]. Hashemi M, Hall M (2019) Detecting and classifying online dark visual propaganda. Image Vis Comput 89:95–105
[32]. Ho, W. H., & Watters, P. A. (2004). Statistical and structural approaches to filtering internet pornography. IEEE International Conference on Systems, Man and Cybernetics. 5, pp. 4792-4798. IEEE.
[33]. Howard, A. G. (2013). Some improvements on deep convolutional neural network based image classification. arXiv preprint , arXiv:1312.5402.

[34]. Hu W, Wu O, Chen Z, Fu Z, Maybank S (2007) Recognition of pornographic web pages by classifying texts and images. IEEE Trans Pattern Anal Mach Intell 29(6):1019–1034
[35]. HuW, Zuo H,Wu O, Chen Y, Zhang Z, Suter D (2011) Recognition of adult images, videos, and web page bags. ACM Transactions on Multimedia Computing, Communications, and Applications 7(1):28
[36]. Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint , arXiv:1508.01991.
[37]. Ioffe, S., & Forsyth, D. (1999a). Finding people by sampling. The Seventh IEEE International Conference on Computer Vision. 2, pp. 1092-1097. IEEE.
[38]. Ioffe, S., & Forsyth, D. A. (1999b). Learning to find pictures of people. Advances in Neural Information Processing Systems. 11, pp. 782-788. MIT Press.
[39]. Ioffe S, Forsyth DA (2001) Probabilistic methods for finding people. Int J Comput Vis 43(1):45–68
[40]. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., et al. (2014). Caffe: convolutional architecture for fast feature embedding. The 22nd ACM International Conference on Multimedia (pp. 675-678). ACM.
[41]. Jiao, F., Gao,W., Duan, L., & Cui, G. (2001). Detecting adult image using multiple features. International Conference on Info-tech and Info-net Proceedings. 3, pp. 378-383. Beijing: IEEE.
[42]. JingHua B, Xian ZX, ZhiXin L, XiaoPing L (2012) Mixture models for web page classification. Phys Procedia 25(1):499–505
[43]. Jones MJ, Rehg JM (2002) Statistical color models with application to skin detection. Int J Comput Vis 46(1):81–96
[44]. Jurafsky D, Martin JH (2014) Speech and language processing. Pearson, London
[45]. Kim S, Zhang B-T (2003) Genetic mining of HTML structures for effective web-document retrieval. Appl Intell 18(3):243–256
[46]. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional
neural networks. Advances in Neural Information Processing Systems, (pp. 1097-1105).
[47]. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. arXiv preprint ,arXiv:1603.01360.
[48]. Lee, J. Y., & Dernoncourt, F. (2016). Sequential short-text classification with recurrent and convolutional neural networks. arXiv preprint , arXiv:1603.03827.
[49]. Lee PY, Hui SC, Fong AC (2002) Neural networks for web content filtering. IEEE Intell Syst 17(5):48–57
[50]. Lee PY, Hui SC, Fong AC (2005) An intelligent categorization engine for bilingual web content filtering. IEEE Transactions on Multimedia 7(6):1183–1190
[51]. Lee J-H, Yeh W-C, Chuang M-C (2015) Web page classification based on a simplified swarm optimization.Appl Math Comput 270(1):13–24
[52]. Li L, Helenius M (2007) Usability evaluation of anti-phishing toolbars. J Comput Virol 3(2):163–184
[53]. Li H, Xu Z, Li T, Sun G, Choo K-KR (2017a) An optimized approach for massive web page classification using entity similarity based on semantic network. Futur Gener Comput Syst 76(1):510–518
[54]. Li X, Rao Y, Xie H, Lau RY, Yin J, Wang FL (2017b) Bootstrapping social emotion classification with semantically rich hybrid neural networks. IEEE Trans Affect Comput 8(4):428–442
[55]. Liang, K. M., Scott, S. D., & Waqas, M. (2004). Detecting pornographic images. Asian Conference on Computer Vision, (pp. 497-502).
[56]. Liparas, D., HaCohen-Kerner, Y., Moumtzidou, A., Vrochidis, S., & Kompatsiaris, I. (2014). News articles classification using random forests and weighted multimodal features. Information Retrieval Facility Conference (pp. 63-75). Springer.
[57]. Liu W, Deng X, Huang G, Fu AY (2006) An antiphishing strategy based on visual similarity assessment. IEEE Internet Comput 10(2):58–65

[58]. Luo Y (2017) Recurrent neural networks for classifying relations in clinical notes. J Biomed Inform 72(1):85–95
[59]. McKenna SJ, Gong S, Raja Y (1998) Modelling facial colour and identity with gaussian mixtures. Pattern Recogn 31(12):1883–1892
[60]. Mikolov, T., Karafiát, M., Burget, L., Černocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. 11th Annual Conference of the International Speech Communication Association,2, p. 3.
[61]. Moustafa,M. (2015). Applying deep learning to classify pornographic images and videos. 7th Pacific-Rim Symposium on Image and Video Technology (p. arXiv:1511.08899). At Auckland, New Zealand: arXiv preprint.
ect. Journal of Big Data 7 (2)