TFIDF and Entropy for Sports, News and Gambling Web Content Classification

Muhammad Dawood, Othman Bin Ibrahim, Aliyu Mohammad Abali

Abstract: ***The exponential increase in online data or information brought up the issue of information security. Gambling web content is one of the greatest harmful res ources that pollutes children's and adolescents ' minds by disguis ing sports and News Web pages. Gambling can be adopted by any gender or age. Several web content-based analytical approaches were proposed to prevent children from accessing these illegal web content. They are introduced to the Internet at an early age. Most approaches are weak to classify web content of high similarity such as Gambling, Sports and News Web pages. In this paper two existing term weighting schemes namely TFIDF and Entropy are used as feature selection process in filtering website. We examine the performance of both techniques via datasets and compare it with the term weighting schemes. The suitability of these term weighting schemes as the selection of features is measured according to the accuracy of the results obtained using the classification program known as Support Vector Machine (SVM). In this paper the performance of TFIDF and Entropy is judged on the basis of Accuracy. Results showed that TFIDF performed better than Entropy. On average, TFIDF obtained 97% and Entropy 91% accuracy. This study is hoped to give other researchers an insight, especially those who would like to work in the same area.***

Index Terms**: *Keywords: URL, Term Weighting Schemes, TFIDF, Entropy, And SVM.***

I. INTRODUCTION

Internet users are growing rapidly as information on web pages has become the main source of knowledge for individuals and organizations. The internet was initially introduced just for communicating the term used for communication was known as inter-networking. With time an excessive amount of information started to be shared and added to the internet in the form of web pages. Knowledge seekers while in the form of individuals or organizations started accessing internet. The information on the internet can be factual or fake. Information available on the internet is huge in amount which may or may not be useful. Over 80% of our young people spend a great deal of their time around the world on the internet because the internet was introduced into their lives at a very young age. The new generation and young people are one of the first to experience problems with excessive internet use. [5].

Especially the youth must be secured from such online activities which will affect their lives, Such as their relationship with their family, friends, community and particularly academics and even their own health too. In health, obesity is the biggest problem, whereby they can't go for outdoor games and always sitting in front of a computer. May be for their desire they can start lying and gradually they can go far like stealing and mugging. Most of the youths can face psychological problems. To secure our youth from all these online addictions some appropriate actions must be taken like web filtering. Web filtering is one of the options we have to secure our youth from online addiction which can be in the form of gambling, sports and entertainment web pages to misguide. This can also lead to material for sexual exploitation and online victimization, including harassment or cyber intimidation and sexual solicitation [6].When their personal information is provided on the internet.

Web pages contain different ways to attract internet users for gambling, like in the name of sports and entertainment, but further redirect them onto websites of gambling and introduce betting to the viewers. And introduce them earning money online. New users or young ones especially get attracted towards these web pages. Most of them think it’s the easiest way to earn money. In the same way the youth can be deceived and addicted to gambling. Excessive internet use is emerging as the most negative aspect of online activities for young people. The Internet presents young people with both risks and opportunities to obtain information or become addicted to some online activities.

With increasing technological advances enabling extensive use of personal computers and easy access to low-cost / high-speed internet, it has become a more popular gambling activity among youth internet gambling [4]. Despite legal attempts to restrict internet gambling, there is an increasing number of youth gambling online. A study of high school and college students in Canada and the U.S., as well as other online respondents, found that 9% of high school students, 6% of college students, and 42% of online respondents have ever played for money on the internet [10]. For money, those who have ever played on the internet have been reported to have played 70 percent on a weekly basis. Who played among them on a weekly basis, 85% were male, 23% were 18-24 years of age, and 28% were problem gamblers. According to the 2010 National Annenberg Youth Risk Survey, the rate of monthly Internet gambling activity among 18-22-year-old American males rose significantly from 4.4% in 2008 to 16% in 2010. The same study also found an increase in internet gambling among high school-age respondents,

Revised Manuscript Received on 04 May 2019

Muhammad Dawood, PhD student in faculty of computing, UTM, Malaysia, mastangalbaloshi@gmail.com

Othman Bin Ibrahim, Associate Professor in faculty of Computing, UTM, Malaysia

Aliyu Mohammad Abali, PhD student in faculty of computing, UTM, Malaysia.

although the change was not statistically significant [12]. These findings suggest that internet gambling may become more popular with youth, particularly among college-age males, and links to problem gambling may be developed.

As shown in recent studies, Internet addiction may affect individuals of any gender, age or socio - economic status. Adolescents may be affected by life, academic performance, social, psychological and physical well - being, and limited by excessive internet use [2]. It seems that the prevalence of gambling among adolescents remains high today. In a national survey of U.S. residents aged 14-21, in the past year nearly 7 out of 10 people had played [16].

II. BACKGROUND AND PREVIOUS WORK

A filtering of web content and internet filtering is defined as blocking unwanted internet content. It is possible to deny and filter access to some unwanted material. A person from the filter set will provide a system to block certain data. It may include advertising, a virus, file transfer, or other material that is violent. To complete the request, a request made by a user to a blocked web page will be maintained with the internet filter. It will be totally blocked or redirected to another location. But to control what content a web content reader is allowed to filter, filtering is designed specifically for that, usually used by organizations like offices and universities to prevent users from accessing and viewing inappropriate content or websites. Web filtering is used in many fields nowadays. As an online file sharing may be blocked in offices, College laboratories may deny access to certain sites and, in particular, parents and guardians may limit children's online activities.

Whether the Web pages are appropriate or not to be browsed can be differentiated with good Web content filtering software. The linguistic analysis and soft computing approaches are required for a better understanding of the semantics of text through computers. There are now several commercial web filtering products on the market, but the techniques used by these require further improvement, which makes them less efficient, especially with the ever-changing web content of today [10]. Some of them are lacking in linguistic analysis, which also affects precision. Blocking Uniform Resource Locator (URL), Internet Content Selection Platform (PICS) is checked and the most commonly used techniques are keyword matching. Web users ' technique of allowing or Denying access to websites by checking the required URL or Internet Protocol (IP) address with database-set URL or IP address lists is known as URL blocking. The main problem with URL blocking is that it is difficult to get a complete up-to-date list of URLs with current limited web technology since it is estimated that 571 web pages are being added per minute and that it becomes 0.8 million web pages per day (Internet 2012). This technique's maintenance and frequency fail when unknown web content becomes expensive. Meanwhile, PICS has always had a problem in checking the trust technique. Thus, due to the inherent weakness of the ever - changing web content, PICS can only be suggested as an additional filtering technique. On the other hand, keyword matching is a technique that blocks web pages based on the occurrence of prohibited keywords on that particular web page. It has been designed to overcome the dynamic content issues, but it is not efficient enough to distinguish Web content from different topics, but similar terminologies [10]. This technique, for example, will block pornography and sex education web pages even with the intention of blocking only pornography web pages. This will enable Web users to access banned Web pages without blocking the system. Furthermore, the system will also prevent access to healthy web pages by web users. Consequently, the content-based technique which includes a deeper understanding of the semantics of text and other Web content items would probably be better suited to prohibiting or illegal Web pages [8].

Recently, Analysis of sentiment has become one of the growing areas of research related to natural language processing and machine learning. Many opinions and sentiments on specific topics are available online, enabling multiple parties such as customers, businesses and even governments to explore these views [1].

III. ENTROPY TERM WEIGHTING SCHEME

The method of entropy is based on probabilism analysis. TFIDF is different from entropy; it is more often the term is mostly in documentation, the term is more important. However, to meet this requirement document has been removed stop words. Two aspects of entropy are the local and the global term weighting, computes weight. This means that once the weight of each word is calculated in the range of 0 to 1 be the local and global weights. These values are normal. The entropy classification scheme has been implemented for Web news [13]. It is as follows:

$$G_i = \frac{1+\sum_{i=1}^{n} \frac{TF_{ij}}{F_i} \log\left(\frac{TF_{ij}}{F_i}+1\right)}{\log N} \quad (1)$$

L_{ii} ith and jth weight of the document in which the total weight of the term is the local time and the universal time G_i total weight of the ith term. Where N denotes is the total number of documents in a collection. i is a variable that corresponding to the ith term where i =1,2,…, $|\beta|$ while j is a variable that corresponding to the jth term document where i =1,2,…,N . The notation x_i represents the total term weight of the ith term in a collection while TF_{ii} refers to the numerical frequency of the i th term occurs in the j th document. On the other hand, in the DF method term based on document frequency are ranked on each term in a document group [13]. DF within the period specified group containing a number of documents are measured by. This is the formula for DF.

Where DF_i is the number of documents that contain the ith term in the collection. DF assumes that words that occur in the provision of documents within a collection are often. Therefore, DF important step, is a group of documents, the ranking for a specific period of a term is based on the

maximum number of documents [13].

IV. TERM WEIGHTING SCHEMES LIMITATIONS

Illegal web pages such as Gambling structures differ from web pages categorized as news, sports, and entertainment. There may be few words and short terms in these web pages. But in some cases entertainment web pages may contain some data on the other hand gambling pages, if joined with news web pages, may contain more words. Therefore, it may be inappropriate to implement the traditional term weighting scheme for illegally classifying websites primarily for gambling. TFIDF assumes that it is related documents if a word is rare in text categorization. However, there are very few words on gambling web pages. Mostly, Unlawful web pages share offensive terms, as they often appear on a web page. The higher the TFIDF document frequency, the more representative that term is [10] in such a case. TFIDF and entropy are weak to identify "grey area "issues.

V. GAMBLING WEBPAGE CLASSIFICATION METHOD

Classification between sports and news has been always a challenge to determine it's healthy, and Gambling as inappropriate webpages especially when it comes as content based. All these webpages are having tremendous amount of text. Gambling webpages have different characteristics from sports and news webpages where the document length for most of them are lengthy having tremendous amount of text as compared to Gambling webpages.

Moreover, it also made it difficult and challenging to classify the web, without any restriction on the word used in web pages. This is due to the high dimensionality of text features, which not only consumes high computation time, but also complicates and complexes the classification process. We compared the following two term weighting schemes known as TFIDF and Entropy to meet the objective of classifying webpages for gambling. Basically, the design of the testing process consists of four main parts which are web page retrieval, pre-processing, feature selection, and classification parts as shown in Fig. 1.

First, Pre-processing as text documents the retrieved web pages. Text documents (Doc) would go through the selection process of the term feature to reduce the original data vectors to a small number of representative term features. Features will be shown as vectors of term weight (E) and treated as input for the classification part support vector machine (SVM). Finally, the result of the SVM classification will be examined under the process of performance assessment.

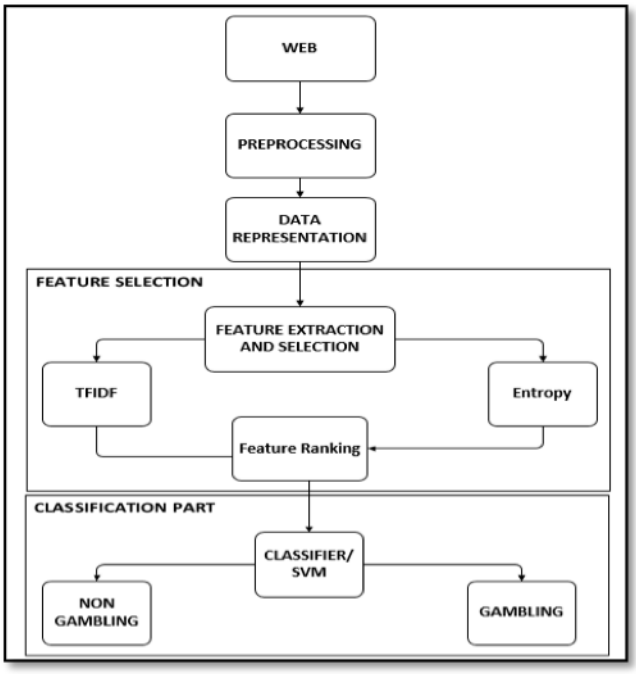

Fig.1: Classification of gambling webpages.

A. Web Data Collection

Collecting Web documents in a database through retrieving and storing is known as Web data collection. For the next phase of the analysis important information is the HTML code, text and image. Crawler can be used to collect data.

B. Pre-processing

In pre-processing of the Web page are transformed into a text or image documents. During the pre-processing phase, extraction of content related to text or image is performed from the Web page. HTML codes are excluded from this process [7]. The documents go through stopping and stemming process which is performed for the text content [13].

C. Data Representation

To transform data into an understandable form of computer like text or image is performed in the data representation process. This text or image form needs further transformation for making a set of numerals which are readable for digital form. The selected text content from any Web page will be converted as an input for Vector Space Model (SVM), in the next section it will be discussed in detail.

D. Feature Extraction and Selection

This process describes that to summarize the data and transform it into a representative set of features that will work as a classifier. On the other hand, the method of selecting a subset of relevant features in machine learning is used to improve the learning models feature selection method. This process will reduce the complexity of data as a set suitable for classifiers.

E. Training and Testing

For testing, training and classification processes various techniques are found for Web content analysis. One of the interesting methods for automatically classifying Web

content analysis on the content base is to introduce a training process. This learning and adaptation process involves supporting vector machine (SVM), ANN and other machine learning techniques. This process will include a large number of examples of testing and training that have both positive and negative elements.

F. Classification

The classification process is the process which deals with the input data to be classified into various groups. On behalf of Web filtering, have to separate forbidden from the useful material through SVM.

VI. SUPPORT VECTOR MACHINE

SVM training is a set of positive examples based on the learning process of a linear hyperplane separation from negative examples. Hyperplane is the closest point to the positive and negative example in hyperspace (called support vectors) [3]. The SVM classification procedure is shown in Fig. 2. Bold line separates the plane representing two examples classes. The other two lines are to inform the nearest positive and negative examples of the hyperplane (support vector). Two linear classifiers elements, a perpendicular weight vector hyperplane (the training set that accounts for feature-representing components) and a bias B that determines the hyperplane offset are based on. An unlabelled example ẍ is classified as positive $f(\ddot{x}) = \dot{w}\ddot{x} + b \geq 0$, Otherwise it will be classified as negative. SVM is thus a binary classifier. Job application as an SVM text classifier includes [14].

Fig.2: Support Vector Machine.

As a SVM text classifier, there are several advantages. First, SVM can handle many features exponentially, Examples in its transformed space do not need to be represented, and only two examples of similarity are needed to be calculated efficiently. Extra features and characteristics of high dimensions are handled correctly; such an aggressive feature selection SVM is required [9]. Second, the process of error estimating formulas which predicts how well a ranking SVM can help in working. This technique eliminated the need for cross-validation. However, there are some drawbacks as well. First of all, both training and test VSM his speed and size limit. In SVM classification process is very time consuming. Second, SVM is only directly applicable to binary classification. And is the order of several binary problems. Third, SVM linearly distributed vacuum. So, not linearly distributed spaces can sometimes hyperplane [11].

For illegal web content filtering SVM is suitable. However, a highly complex feature will be representing the web content. While SVM does not require aggressive selection of features, its complexity results in high resource and time consumption in training and testing.

VII. EXPERIMENT AND RESULT

A. Data

In order to make standard classification, we collected 25 web pages from each category. This study will classify legal (non-gambling) and illegal (gambling) web pages into two categories. Illegal web pages refer to web pages showing gambling activity. The non-gambling web pages, on the other hand, refer to those web pages that display news and sports content web pages. However, the dataset for non-gambling category that will be used here contains web pages for Sport and News. Table I summarizes the information.

Table I: description of Web pages as Dataset label

S/No	Category	Label	Web pages
1	Gambling	Gamb	25
2	News	New	25
3	Sports	Spo	25
	Total		75

The dataset is divided into three data types shown in Table II. Extensive examination and identification of the performance of the weighting scheme for each term. Dataset 1 is used to test the ability of term weighting schemes to properly classify "gray area" news web pages. The gray area condition is the condition that the classifier would be confused by web pages containing high similar terminology for several topics. For example, if we look at the terminology used in the web pages of Gambling, Sports and News such as' play, game, gaming,' rules' and so on all of these web pages, whether gambling or non-gambling. As a result, it is difficult for most classifiers to distinguish between web pages with similar terminology. Dataset 2 is designed to test the capacity of term weighting schemes to classify Sports web pages correctly. Dataset 3 will represent the actual situation of the internet, where all categories of Web pages i.e. gambling, sports and news will be assorted together in the real internet environment. The experimental results will be presented in Section 7.3.

B. Experiment Environment

We implemented the Support Vector Machine as our classifier to perform classification. As shown in Table II, we used three types of datasets. Datasets will be divided into 3 sets. Each dataset will be used

to determine the performance of selection criteria under 3 conditions which are gambling, sports and news web pages. Dataset 1 is used for the examination of the selection criteria for Web pages related to news. Dataset 2 will be used to examine selection criteria under Web pages of sports. Finally, dataset 3 will represent the actual situation of the internet, where all categories of Web pages i.e. gambling, sports and news will be assorted together in the real internet environment. Our experiment includes a feature selection process and classification process, for feature selection further two Term weighting schemes as TFIDF and Entropy will be used additionally for classification SVM will categorize and assign terms to the groups. Furthermore, it covers the evaluation process to evaluate the performance.

C. Results and Discussion

The experiment was carried out by testing schemes for TFIDF and Entropy term weighting with three categories of data sets shown primarily in Table II. In this experiment, we selected our terms differently for standard evaluation purposes, we introduced terms with number of repetition in the datasets. Let me explain the word repetition for terms here; because of the abundant content available in News and Sports related web pages some gambling web pages also, we were attracted to select our terms on the basis or repetition in the web page. The repetition of terms 10 times, 30 times and 50 times respectively in each category. Finally, to judge their effectiveness in both term weighting schemes, we measured using precise, accurate and F1 standard information recovery measurements.

Table II. Categories of datasets for Testing and Training

Data	Category	Training	Testing
Dataset 1	Gamb (25)	15	**10**
	New (25)	15	**10**
Total	50	30	**20**
Dataset 2	Gamb (25)	15	**10**
	Spo (25)	15	**10**
Total	50	30	**20**
Dataset 3	Gamb (25)	15	**10**
	New (25)	15	**10**
	Spo (25)	15	**10**
Overall	75	45	**30**

D. Experiment on Dataset 1

The experiments were conducted using dataset 1, 2 and 3; as explained in section 7.1 and 7.2. Fig.3 describes the experimental results that were conducted using dataset 1. Standard measurements of retrieval of information such as accuracy, recall, F1 and accuracy were used to assess the performance of term weighting schemes as methods of selection of features. To examine each term weighting scheme's performance by selecting different number of features, each increment of 25 features is taken as a benchmark evaluation. The data is collected from the web pages of Gambling and News in dataset 1. In this dataset, the terms in News Web pages are higher than in Gambling.

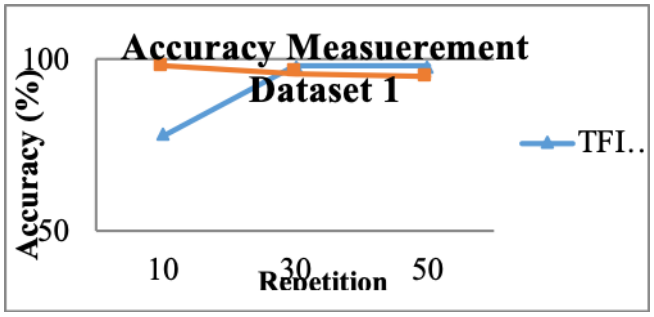

Fig.3: Accuracy Dataset 1.

TFIDF has achieved the highest value in Dataset 1. But here it is noticed that the increase of features is increasing the accuracy of both of the term weighting schemes, which means the more the number of features the more the frequency of word or terms appear.

E. Experiment on Dataset 2

Dataset 2 is another setup to select the most important features in dataset 2; data is collected from the web pages of Gambling and Sports. In this dataset the terms are greater in number in Sports Web pages as compared to Gambling. Similar to previous experiments, several experiments based on TFIDF and Entropy term weighting schemes have been conducted using dataset 2.

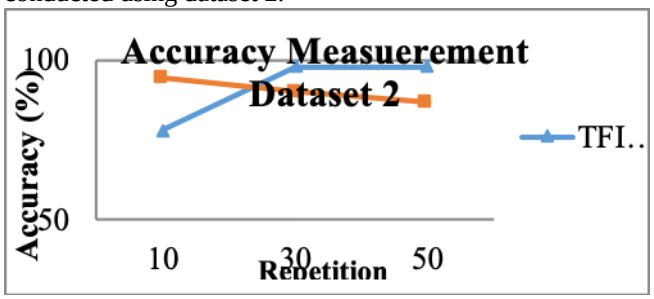

Fig.4: Accuracy Dataset 2.

TFIDF has achieved the highest value with every number of features. With the increase in the features gain of accuracy in both of the term weighting schemes, which means the more the number of features the more the frequency of words appears.

F. Experiment on Dataset 3

In dataset 3, the data is collected from Gambling, News and Sports Web pages. In this dataset the terms are greater in number in Sports and News Web pages as compared to Gambling.

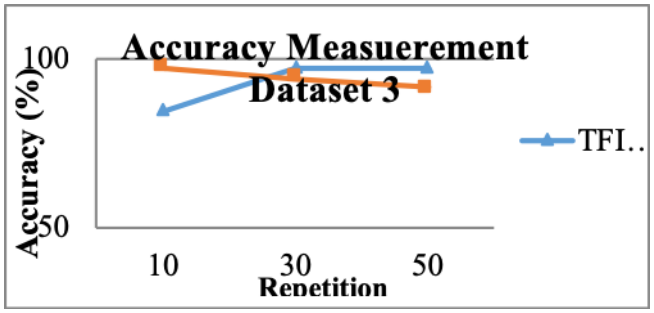

Fig.5: Accuracy Dataset 3.

TFIDF has achieved the highest value in all Datasets. The result might fluctuate if

many features or data used in testing and training in the classification process. This result is with only a range of 1-75 features, and the result showed the accuracy increased with the feature increase. Whereas Entropy could not gain more accuracy than TFIDF, may be due to the number of unique terms. Since the number of unique words or terms in the collection set is large, PCA (main component analysis) is required to select the most relevant classification feature [13]. If the human intervention is considered in Entropy then the results showed better performance by Entropy. But when the document length gets bigger the performance decreases.

VIII. CONCLUSION

The assumption was made that one of the important factors affecting the classification of the Gambling page is the length of the Web page. Several experiments were conducted and the results were obtained to distinguish the performance of term weighting schemes TFIDF and Entropy. And the highest value in all Datasets is clearly shown by TFIDF. This study found the feature pattern to be the primary factor influencing the performance of classification. The more the classifier captures the feature, the better the classification of the web. The Web content filtering approach could be further improved by including multilingual analyses can be included in the feature selection process in order to enable gambling Web filtering approach to handle a broader selection of languages. There is a possibility of the content analysis that include more Web topics by identifying the term feature patterns of each topic.

ACKNOWLEDGEMENT

I would like to thank Assoc. Prof. Dr. Othman Bin Ibrahim for his guidance and support. The knowledge I gained was invaluable and the challenges we faced were a very enriching experience for my career.

REFERENCES

[1] Alotaibi SS. Sentiment Analysis in the Arabic Language Using Machine Learning: Colorado State University. Libraries; 2015.
[2] Beard, K. W.andWolf, E. M. (2001). Modification in the proposed diagnostic criteria for Internet addiction. CyberPsychology & Behavior, 4(3), 377-383.
[3] Boser, B. E., Guyon, I. M.andVapnik, V. N. (1992). A training algorithm for optimal margin classifiers. Paper presented at the Proceedings of the fifth annual workshop on Computational learning theory.
[4] Derevensky, J. L.andGupta, R. (2007). Internet gambling amongst adolescents: A growing concern. International Journal of Mental Health and Addiction, 5(2), 93-101.
[5] DiNicola, M. D. (2004). Pathological Internet Use Among College Students: The Prevalance of Pathological Internet Use and Its Correlates. Ohio University.
[6] Guan, S. S. A.andSubrahmanyam, K. (2009). Youth Internet use: risks and Opportunities. Current opinion in psychiatry, 22(4), 351-356.
[7] Gupta, S., Kaiser, G., Neistadt, D.andGrimm, P. (2003). DOM-based content extraction of HTML documents. Paper presented at the Proceedings of the 12th international conference on World Wide Web.
[8] Hammami, M., Chahir, Y.andChen, L. (2006). Webguard: A web filtering engine combining textual, structural, and visual content-based analysis.Knowledge and Data Engineering, IEEE Transactions on, 18(1), 272-284.
[9] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. Machine learning: ECML-98, 137-142.
[10] Lee, Z. S., Maarof, M. A., Selamat, A.andShamsuddin, S. M. (2008). Enhance Term Weighting Algorithm as Feature Selection Technique for Illicit Web Content Classification. Paper presented at the Intelligent Systems Design and Applications, 2008. ISDA'08. Eighth International Conference on.
[11] Navigli, R. (2009). Word sense disambiguation: A survey. ACM computing surveys (CSUR), 41(2), 10.
[12] Romer, D. (2010). Adolescent risk taking, impulsivity, and brain development: Implications for prevention. Developmental psychobiology, 52(3), 263-276.
[13] Selamat, A.andOmatu, S. (2004). Web page feature selection and classification using neural networks. Information Sciences, 158, 69-88
[14] Soderland, S. (1997). Learning to extract text-based information from the world wide web. Paper presented at the Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining.
[15] Tong, S.andKoller, D. (2002). Support vector machine active learning with applications to text classification. The Journal of Machine Learning Research, 2, 45-66.
[16] Welte, J. W., Barnes, G. M., Tidwell, M. C. O.andHoffman, J. H. (2008). The prevalence of problem gambling among US adolescents and young adults: Results from a national survey. Journal of Gambling Studies, 24(2), 119-133.