Classification of Competition Content on Web Pages using Machine Learning Algorithm

1st Dinda Nabila Amartha  
*School of Applied Science  
Telkom University*Bandung, Indonesia   
dnabila0703@gmail.com2nd Dedy Rahman Wijaya   
*School of Applied Science   
Telkom University*Bandung, Indonesia  
dedyrw@telkomuniversity.ac.id3rd Suryatiningsih  
*School of Applied Science   
Telkom University*Bandung, Indonesia  
suryatiningsih@telkomuniversity.ac.id

*Abstract*— To develop quality and human resources in the future, students as an agent of change can increase their potential by participating in competitions. There are so many benefits that can be obtained from participating in competitions such as increasing abilities, improving soft skills, increasing experience, expanding relationships, and of course, getting prizes and certificates. Competition information can be found on the internet, especially in search engines. But, even though it has been optimized by SEO, it is not uncommon to find web pages that provide invalid competition information and pass the deadline. So a web pages classification is needed. This paper uses a machine-learning algorithm to classify valid competition information web pages for students with the Gradient Tree Boosting classifier method. Then, we got 85% accuracy. Of course, this is a good model to classify and validate the competition information web pages effectively and efficiently. Implementation of the invention will be discussed.

Keywords—Competition, Student, Web pages, Machine Learning, Gradient Tree Boosting

# Introduction

To maximize and develop the potential that students have, they must be able to take advantage of opportunities well, one of which is participating in competitions. There are so many benefits that can be obtained from participating in competitions such as increasing abilities, improving soft skills, increasing experience, expanding relationships, and of course, getting prizes and certificates. The internet provides a lot of competitive information, one of which is by using a search engine. A fundamental definition of search engine speaks of information retrieval based on "keyword".

Many factors affect a site's ranking and web pages’ validity in search engines. Each result that is calculated in the google search engine is determined by a keyword [1].

Diagram

Description automatically generated

Fig. 1. Indexing Process

The effectiveness and accuracy of a web page depend on its ranking in search engines and other factors including keywords. Therefore Search Engine Optimization (SEO) has the role of knowing whether certain web pages have been optimized for certain keywords according to the guidelines of popular search engines [2]. To describe the content that users are looking for, SEO provides accurate results suggestions that hint at user needs. Ensuring this gives the best possible outcome between investigation and data [3]. To get the top ranking, the web pages you are looking for need to be indexed with a web crawler. Search engines require active participation from users by creating an index, transparent process, and responding to inquiries [4]. Fig. 1 is the process of an indexing web page, where the process steps are based on number order:

1. To browse the web and retrieve HTML pages, a web crawler is used.
2. Web pages are stored in their original form in search engine’s document databases by a web crawler.
3. Next step is the page undergoing changes made by search engines to extract textual content and information from links for indexing.
4. Then search engines create an index using a single word or phrase to generate direct and replacement page indications. It then logs information about link activity and generates a snippet.
5. And finally search engine stores the results of indexes in the index database.

Sometimes in search engines even though they have been optimized with SEO, sites that provide competition information are still found that are invalid and have passed the time limit. So we need web page classification to determine web validity. There are several reasons why web classification is needed including, manually classifying information on the internet makes it difficult for experts, web pages are dynamic and unstable so a lot of effort is needed for regular classification and more time and effort for manual classification [5].

In the machine learning paradigm, classifiers are built automatically by mapping pre-classified input examples with certain characteristics to the output. The advantages of this approach are web page verification automation, excellent effectiveness, and domain independence [5]. To obtain information on competitions that will be given to students, they usually visit the web pages one by one and cannot decide whether the web pages are valid or not because sometimes there are deadlines that have passed and the content doesn’t match. According to the explanation above, this study proposes a machine-learning algorithm to classify web pages as media information often provide-invalid information about the competition.

The rest of this paper is organized as follows: Section II explains related works. Section III describes the methods and dataset. Implementation results and discussion are shown in Section IV. Section V contains the conclusion and final result.

# related works

After viewing the data sometimes can’t interpret the information that has been extracted in the data. Machine learning is used to teach machines how to handle data more efficiently. Machine learning is the study of how machines learn from experience. Many scientists say "machine learning" is related to the term "artificial intelligence" Intelligence here in a broad sense has the broadest meaning [6]. Machine learning is improving due to a large number of data sets available on the internet [7]. The information on the web page is divided into two layers namely, a warning table or a list of downloadable files then some other details like post date, title, name, then the second layer contains a detailed description [8].

In research, the application of machine learning and data mining methods is needed to turn the information available in the biosciences into valuable knowledge [6]. Many algorithms are applied in machine learning. Gradient tree boosting is one of them by using a tree ensemble technique to show value in this case it is applied to detect rice quality[9]. Feature selection algorithm used to eliminate irrelevant and redundant features [10]. As in a study entitled "Estimating city-level poverty rate based on e-commerce data with machine learning" there, they use feature selection to see poverty levels. The results show that the selected features according to the aspects and items reflect the real conditions in Indonesia [11].

Efficient automated web page classifiers are needed to help validate the information that evolves from the WWW [12]. In the application of web page classification deep learning algorithms study the desired features deeper while conventional machine learning only extracts the desired features from web pages for classification [13]. Text classification is the task of automatically categorizing group documents into one or more classes. Its main purpose is to help users extract information from textual sources [14].

Researching entitled "SulSite-GTB: identification of protein S-sulfenylation sites by fusing multiple feature information and gradient tree boosting" use a subset of features in gradient tree boosting classification to predict S-sulfenylation sites and fivefold cross-validation to evaluate model prediction performance used. And get the results of prediction accuracy of 92.86% and 88.53%, respectively, and AUC values ​​of 0.9706 and 0.9425 in the testing and training set, respectively [15]. To improve system performance, especially for multiclass classification in the future, machine learning algorithms are very much needed [16].

# research methods

## Proposed Method

Fig. 2 shows processes of retrieving URL and data from search engines using python with libraries: request, urllib, pandas, and request\_html. Then after the results are obtained in the form of raw data from the URL, title and content extracted into a dataset that is ready for use in machine learning. The following is a further step-by-step explanation:

Diagram

Description automatically generated

Fig. 2. Proposed Method

### Crawling content from google search engine

Crawling is the process by which the Google search engine searches and scans data on web pages [17]. To get data from search engines in this study using python with libraries: request, urllib, pandas, and request\_html. In this case, the data needed is in the form of links, titles, meta descriptions, date or time, and content. First, create a function to format and encode keyword URLs, then submit them to search engines and show the results. Next, parse the response HTML and then wrap the function in a search engine function, which will bring all of the above together. There are several keywords used such as student competitions, student competitions, and 2022 student competitions. Then indexing is the process by which search engines archive information on the site. And finally, the data was obtained in the form of links, titles, and phrases which are then processed into the required data.

### Extracting Dataset

The results data set still cannot be used for machine learning because the data is textual. Then the data will be extracted into numeric. From the results of the extraction earlier, the features that will be used in machine learning will then be made. Then count the number of words and the length of the text in the title and meta description using the counter function in python. To validate the data from the web pages, labeling is done manually one by one with the number symbol 1 for valid and 0 for invalid. The criteria for determining the validity of web pages that provides information from the competition is seen from the deadline date and the competition content on the web page.

### Normalization

The purpose of normalization is to minimize Inform errors. Because in the application of machine learning the data set is multidimensional. The normalization method applied is the standard scaler. Below is the formula for the standard scaler.

Standard Scaler = (1)

Description :

= Value of feature

= training sample

= standard deviation

### Feature Selection

In the preprocessing process, feature selection is carried out to select features that have an effect and discard features that have no effect. Some relevant features can be found by finding correlations between features and class labels. On the other hand, the high correlation between features will lead to features redundancy [18]. This study used a chi-square test. The formula for the chi-square used is:

(2)

Oi = Observed Value

Ei = Expected Value

### Grid Search

In the grid search parameters will be tested and produce the best parameters. This is very important to classify web pages that provide valid competition information with precise and accurate predictions. Table 2 shows the parameters used in several algorithms in this study. Then each parameter will be tested according to the algorithm to produce the best parameters.

Table 2 Parameters used

|  |  |
| --- | --- |
| **Parameter** | **Value** |
| n\_estimators | 50 |
| 100 |
| 150 |
| 200 |
| learning\_rate | 0.001 |
| 0.01 |
| 0.1 |
| Max\_depth | 1 |
| 3 |
| 5 |
| 7 |
| 9 |

### Ensemble Classifier

The gradient tree boosting method is an ensemble of decision trees method to predict a target label sequentially [19]. The workings of this gradient tree boosting algorithm are to build a decision tree structure and then each decision tree before it. Errors that occur in the first tree will be corrected in the next three and so on until how many trees will be built. Then the evaluation will reduce the problems that have occurred [9]. The data is divided into two parts, namely training and testing and in this study, the parameters are learning\_rate=0.01, n\_estimators=150. In this case, it is used to validate the competition information web pages on search engines.

### Evaluation

To develop machine learning the evaluation stage is needed to determine the accuracy value obtained. Some analytical matrices that can be used to see model performance are accuracy, recall, precision, and F1 score. Then it is displayed into a confusion matrix to explain in detail the amount of data that is classified with a true or false statement on Fig. 4.

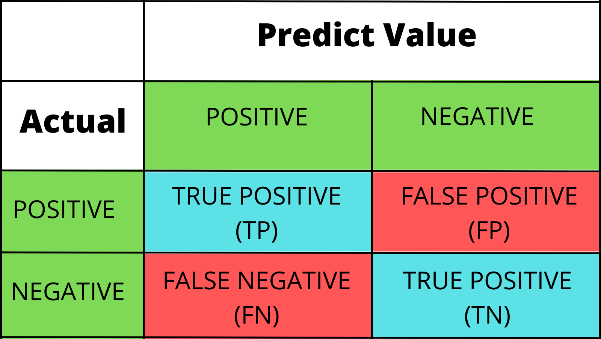


Fig. 4. Confusion Matrix

Four terms represent the results of the classification process in the confusion matrix, namely True Positive, True Negative, False Positive, and False Negative. Let's understand what True Positive, False Positive, False Negative, and True Negative [20]. Below is how the formula to calculate accuracy.

Accuracy = (3)

Precision = (4)

Recall = (5)

F1 Score = (6)

## Dataset

The dataset is obtained from the results of crawling and indexing data from the Google search engine as described above. The data obtained are in the form of links, meta descriptions, content, dates, and titles of each web page visited by entering different keywords. The extracted dataset consists of 14 feature columns and 200 rows. Table 1 is an explanation of the results of the dataset extraction which was previously in the form of text and converted into numeric to be ready to be used in machine learning.

Table 1 Dataset extraction result

|  |  |
| --- | --- |
| Column name | Description |
| Word\_count | Count the number of words in the title and meta description |
| Text\_leng | Count the number of alphabets in the title and meta description |
| Count\_lomba | Count how many “lomba” words are in the title, link, meta description, and content in the link |
| Count\_mahasiswa | Count how many “mahasiswa” words are in the title, link, meta description, and content in the link |
| Count\_deadline | Count how many “deadline” words are in the title, link, meta description, and content in the link |
| Count\_kompetisi | Count how many “kompetisi” words are in the title, link, meta description, and content in the link |
| Count\_pendaftaran | Count how many “pendaftaran” words are in the title, link, meta description, and content in the link |
| Count\_pendaftaran | Count how many “timeline” words are in the title, link, meta description, and content in the link |
| Due\_date | Check the date has passed or not |

Identify the applicable funding agency here. If none, delete this text box.

# results and discussion

## Feature Selection

Table 3 shows the results of feature selection using chi-square and gets the feature ranking results for the feature that ranks first is count\_deadline title and the feature with the lowest ranking is title\_len2.

Table 3 Ranking of Feature

|  |  |
| --- | --- |
| **Rank** | **Feature** |
| 1 | count\_deadline\_title |
| 2 | tanggal\_tenggat |
| 3 | count\_lomba\_content |
| 4 | count\_lomba\_title |
| 2 | count\_deadline\_content |
| 5 | count\_kompetisi\_content |
| 6 | title\_len1 |
| 7 | count\_mahasiswa\_title |
| 8 | word\_count\_title |
| 9 | kata\_timeline |
| 10 | count\_kompetisi\_title |
| 11 | kata\_pendaftaran |
| 12 | count\_mahasiswa\_content |
| 13 | word\_count\_text |
| 14 | title\_len2 |

## Parameters

Table 4 shows the results of the testing parameters in the classification model using a gradient tree boosting classifier. There are learning\_rate,max\_depth, and n\_estimators used.

Table 4 Result of parameters testing

|  |  |
| --- | --- |
| **Parameters** | **Accuracy** |
| learning\_rate= 0.01, max\_depth= 1, n\_estimators=250 | 81% |
| learning\_rate= 0.01, max\_depth= 3, n\_estimators = 100 | 83% |
| learning\_rate= 0.01, n\_estimators = 150 | 85% |

We tried several experiments on the parameters used and the results for the best performance, parameters learning\_rate= 0.01, n\_estimators= 150 were used to classify the validity of the competition information web pages.

## Confusion Matrix

The confusion matrix was used to assess the effectiveness of the classification model. This study uses a gradient tree increasing the valid and invalid classification data, the web page competition information. Then a comparison table will be visualized between the predicted class and the current class.

Fig. 5 shows the web pages classification results from gradient tree boosting. 27 web pages are declared invalid, 4 are declared valid or not yet classified, then 24 competition information web pages are valid and 5 are wrong.

Square

Description automatically generated

Fig. 5. Confusion Matrix

## Classification Report

To measure the predictive quality of the classification algorithm, a classification report is needed. More specifically, True Positives, False Positives, True Negatives, and False Negatives are used to predict the metrics of a classification report as shown in Table 5.

Table 5 Classification Report

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** |
| 0.0 | 84% | 87% | 86% |
| 1.0 | 86% | 83% | 84% |
|  |  |  |  |
| Accuracy |  |  | 85% |
| Macro avg | 85% | 85% | 85% |
| Weighted avg | 85% | 85% | 85% |

From the explanation of the previous confusion matrix, the results of the calculation are as follows, the precision of a valid prediction is 84%, then the recall of the positive cases is 87%, and then on the F1 - Score competition web pages which are correctly predicted positive is 85%.

# conclusion

This research result shows how a machine learning algorithm, namely gradient tree boosting, can classify web pages that provide information about competitions for students, with an accuracy of 85% and the parameters used are learning\_rate = 0.01, n\_estimators= 150. This proves that this method is a good classification because it is more effective and efficient to retrieve valid competition information on the web pages. In future work, we plan to modify or improve the gradient tree boosting classifier method to improve accuracy.

##### References

[1] H. Banaei and A. R. Honarvar, “Web Page Rank Estimation in Search Engine based on SEO Parameters using Machine Learning Techniques,” *Int. J. Comput. Sci. Netw. Secur.*, vol. 17, no. 5, pp. 95–100, 2017.

[2] G. Matoševi, J. Dobša, and D. Mladeni, “future internet Using Machine Learning for Web Page Classification in Search Engine Optimization,” 2021, doi: 10.3390/fi13010009.

[3] S. Chhabra, R. Mittal, and D. Sarkar, “Inducing factors for search engine optimization techniques: A comparative analysis,” *India Int. Conf. Inf. Process. IICIP 2016 - Proc.*, 2017, doi: 10.1109/IICIP.2016.7975341.

[4] C. Practices, “Understanding Search-Engine Optimization,” no. October, pp. 43–52, 2015.

[5] V. D. K. Kasula, “Performance analysis of layered architecture to integrate mobile devices and grid computing with a resource scheduling algorithm,” *Proc. - Int. Conf. Comput. Intell. Multimed. Appl. ICCIMA 2007*, vol. 4, pp. 352–356, 2008, doi: 10.1109/ICCIMA.2007.342.

[6] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, “Machine Learning and Data Mining Methods in Diabetes Research,” *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 104–116, 2017, doi: 10.1016/j.csbj.2016.12.005.

[7] M. Batta, “Machine Learning Algorithms - A Review ,” *Int. J. Sci. Res. (IJ*, vol. 9, no. 1, pp. 381-undefined, 2020, doi: 10.21275/ART20203995.

[8] P. San Ang *et al.*, “Augmenting Product Defect Surveillance Through Web Crawling and Machine Learning in Singapore,” *Drug Saf.*, vol. 44, p. 44, 2021, doi: 10.1007/s40264-021-01084-w.

[9] I. Aulia, D. R. Wijaya, and W. Hidayat, “Rice Quality Detection Using Gradient Tree Boosting Based on Electronic Nose Dataset,” *AIMS 2021 - Int. Conf. Artif. Intell. Mechatronics Syst.*, 2021, doi: 10.1109/AIMS52415.2021.9466073.

[10] D. R. Wijaya and F. Afianti, “Stability Assessment of Feature Selection Algorithms on Homogeneous Datasets: A Study for Sensor Array Optimization Problem,” *IEEE Access*, vol. 8, pp. 33944–33953, 2020, doi: 10.1109/ACCESS.2020.2974982.

[11] D. R. Wijaya, N. L. P. S. P. Paramita, A. Uluwiyah, M. Rheza, A. Zahara, and D. R. Puspita, “Estimating city-level poverty rate based on e-commerce data with machine learning,” *Electron. Commer. Res.*, 2020, doi: 10.1007/s10660-020-09424-1.

[12] C. M. Chen, H. M. Lee, and Y. J. Chang, “Two novel feature selection approaches for web page classification,” *Expert Syst. Appl.*, vol. 36, no. 1, pp. 260–272, Jan. 2009, doi: 10.1016/J.ESWA.2007.09.008.

[13] A. Gupta and R. Bhatia, “Ensemble approach for web page classification,” *Multimed. Tools Appl.*, vol. 80, no. 16, pp. 25219–25240, 2021, doi: 10.1007/s11042-021-10891-3.

[14] A. K. Sharma, V. Shrivastava, and H. Singh, “Experimental performance analysis of web crawlers using single and Multi-Threaded web crawling and indexing algorithm for the application of smart web contents,” *Mater. Today Proc.*, vol. 37, no. Part 2, pp. 1403–1408, 2020, doi: 10.1016/j.matpr.2020.06.596.

[15] M. Wang, X. Cui, B. Yu, C. Chen, Q. Ma, and H. Zhou, “SulSite-GTB: identification of protein S-sulfenylation sites by fusing multiple feature information and gradient tree boosting,” *Neural Comput. Appl.*, vol. 32, no. 17, pp. 13843–13862, 2020, doi: 10.1007/s00521-020-04792-z.

[16] D. R. Wijaya, R. Sarno, E. Zulaika, and S. I. Sabila, “Development of mobile electronic nose for beef quality monitoring,” *Procedia Comput. Sci.*, vol. 124, pp. 728–735, 2017, doi: 10.1016/j.procs.2017.12.211.

[17] J. Kim, S. Youm, and Y. Shan, “Analysis of Fire Accident Factors on Construction Sites Using Web Crawling and Deep Learning Approach,” p. 13, 2021, doi: 10.3390/su132111694.

[18] D. R. Wijaya, R. Sarno, and E. Zulaika, “Sensor array optimization for mobile electronic nose: Wavelet transform and filter based feature selection approach,” *Int. Rev. Comput. Softw.*, vol. 11, no. 8, pp. 659–671, 2016, doi: 10.15866/irecos.v11i8.9425.

[19] Q. Li, Z. Wu, Z. Wen, and B. He, “Privacy-preserving gradient boosting decision trees,” *AAAI 2020 - 34th AAAI Conf. Artif. Intell.*, pp. 784–791, 2020, doi: 10.1609/aaai.v34i01.5422.

[20] “Confusion Matrix.” https://socs.binus.ac.id/2020/11/01/confusion-matrix/ (accessed Feb. 17, 2022).