Quality Classification of ASEAN Wikipedia Articles using Statistical Features

Kanchana Saengthongpattana
Language and Semantic Technology Laboratory
National Electronics and Computer Technology Center (NECTEC)
Pathumthani, Thailand
kanchana.sae@nectec.or.th

Thepchai Supnithi
Language and Semantic Technology Laboratory
National Electronics and Computer Technology Center (NECTEC)
Pathumthani, Thailand
thepchai.supnithi@nectec.or.th

Nuanwan Soonthornphisaj
Department of Computer Science
Faculty of Science
Kasetsart University
Bangkok, Thailand
fscinws@ku.ac.th

***Abstract*— The quality of Wikipedia articles is still the main concerned in all languages. Wikipedia relies mostly on human editors and administrators to provide the quality of content. But the magnitude of Wikipedia content makes locating all instances of article very time consuming. Therefore, we need the automatic quality detection that can help users to evaluate the quality of articles. In this paper, we propose the feature set to applied for the ASEAN language Wikipedia articles. We investigate the statistical features such as # of link, # of infobox, length of article, # of headings, # of files, # of contributors, # of viewer, # of written articles found in other languages , and # of templates applied in the article. The experiments are perform using Naïve Bayes and Decision tree algorithm. We found that the accuracy of Decision tree (96.34%) outperform Naïve Bayes (86.47%). Moreover, we found that the statistical features play an important role in quality classification of Vietnamese, Indonesian, Malaysian, Thai, and Tagalog/Philippines Wikipedia articles.**

Keywords—Quality of articles, Southeast Asian languages Wikipedia, Naïve Bayes, Decision tree, Statistical feature

I. INTRODUCTION

Wikipedia is considered as a rich source of information to everybody. It is widely used as reference source in many documents. Currently, there are 301 written languages found in Wikipedia articles including Southeast Asian languages. The number of English Wikipedia is 5,709,509 articles where as the number of Southeast Asian language is 2,198,529 articles. The content is available in multiple language versions, for example "Chiang Mai Province" in Thai(th), Vietnamese(vi), Indonesian(id), Malaysian(ms), Tagalog/Philippines(tl), and Laos(lo) ;"จังหวัดเชียงใหม่" "Chiang Mai (tỉnh)","Provinsi Chiang Mai", "Chiang Mai (wilayah)", "Lalawigan ng Chiang Mai", and "ຈ້ງຫວັດຊຽງໃຫມ່".

Since the framework of Wikipedia allows user to create or edit articles, therefore we need the classifier to automatically do quality classification. We found that Wiki Project set up the quality assessment by grading the quality level as A, GA, B, C, Start, Stub, FL and List [1]. However, these predefined levels are not appropriate for some language version since there are fewer number of articles and the writing style is different compared to those of English articles.

Consider articles written in Southeast Asian language, we found that there are quite a few number of quality label given by domain experts. That means the quality of a large number of articles have not yet been evaluated. The Thai Wikipedia assigns quality level that includes featured articles and good articles. As of August 20, 2018 only 287 articles out of a total of 125,936 articles on the Thai Wikipedia are labeled as the featured articles and good article. Other Southeast Asian language articles have various number of high and low quality articles (see TABLE II for detail)

As we know that Wikipedia relies mostly on human editors and administrators to contribute to the quality of content. However the magnitude of Wikipedia content makes quality labelling for all instances of article very time consuming. Therefore many researches try to construct automatic method base on classification or clustering techniques to solve the problem. We know that, the success learning algorithm needs informative features to determine the quality of articles.

In this paper, we propose a set of statistical features and study on the performance of machine learning algorithms which are decision tree and Naïve Bayes in order to predict the quality of Vietnamese, Indonesian, Malaysian, Thai, and Tagalog/Philippines Wikipedia articles.

II. RELATED WORK

Researches related to Wikipedia has been widely investigated. Many approaches are studied to explore the different feature sets in order to classify the quality of articles. For example, meta data and textual features [2], length of content, number of external link and number of image [3], Style and variety of words [4] and templates that users add to the article [5]. Contribution of Wikipedia reviewers [6] and the characteristics of editors [7],[8] are studied to predict the quality of Wikipedia articles. However, we found that most these researches focus on the majority languages such as English, German, Spanish.

For ASEAN language, many statistical features of Thai Wikipedia articles are deployed to cluster the quality of articles [9] and the classification framework considers the quality of article in term of its comprehensive content [10]. Japanese Wikipedia is used as resources to build a large scale and general purpose Japanese ontology through ontology learning [11].

In this paper, we decided to focus primarily on those aspects that can help improve the quality of the article in the

Asian languages so we consider the content of the article and its metadata.

III. DATA SET AND FEATURE EXTRACTION

A. *The number of articles in ASEAN languages*

Most of researches on the quality model of Wikipedia articles is focused on the "largest" language – English. In this paper we consider the top 5 of ASEAN languages (Vietnamese, Indonesian, Malaysian, Thai, Tagalo). Note that articles written by Singaporean and Brunei people use the same language as Malaysia which are "Bahasa Malaysia". Tagalog is the written language of Filipinos.

Table 1 shows the statistic of number of articles, number of edits articles, number of user and active users for 6 months. The column name "Parallel Thai", shows the numbers of articles in other languages that match Thai language articles, The top is Vietnamese (68,652 articles). The list is sorted by number of articles compared to all languages in Wikipedia. Vietnamese is No. 12. (number of article is 1,185,557), Thai Wikipedia is No. 56. (number of article is 125,936). The statistic of English Wikipedia and ASEAN languages Wikipedia.

TABLE I. THE STATISTIC OF WIKIPEDIA VERSIONS IN ENGLISH AND ASIAN LANGUAGES

No	Wiki	# of Articles	# of Edits	# of Users	# of Active Users	#of Parallel Thai
1	English	5,700,767	850,975,428	34,280,172	120,767	93,985
12	Vietnamese	1,185,557	41,838,077	615,439	1,687	68,652
24	Indonesian	437,663	14,093,305	985,579	2,493	63,542
29	Malaysian	318,907	4,399,051	219,655	442	38,172
56	Thai	125,936	7,772,754	327,082	1,189	125,936
68	Tagalog	81,633	1,647,833	94,562	124	24,436
95	Myanmar	39,311	417,451	52,718	95	7,052
159	Khmer	6,654	215,813	24,605	81	3,184
209	Loas	2,868	65,642	10,406	17	3,358

B. *The quality classes of articles*

There are 7 quality classes for English content of Wikipedia varied from the highest to the lowest quality; Featured Article (FA), Good Article (GA), A-class, B-class, C-class, Start, Stub. Other languages article applies less grading scale such as Polish article has 5 quality classes. In Wikipedia there is no common standard classification of quality articles among different language articles [14]. Some languages applies expanded rating scale (EN, RU), others are limited to 2-3 grades (BE, DE). In other words, each language version can have its own classification system of articles quality, but all of them use at least two highest classes - equivalent for FA and GA. Vietnamese, Indonesian, Malaysian, Thai, and Filipino Wikipedia articles use two highest classes (Featured and Good Article). We consider the Featured and Good article tags as the "High Quality Class". The high quality means that the Wikipedia community agreed with the content quality by providing the remarkable symbol, the star, and plus sign to notify all users. However the low quality articles can be assured by Broom symbol which means that the content of article need to be rewrited, categorized, put link, or it is too short. The Broom symbol not provided in Wikipedia articles of Vietnam. Therefore we used Puzzle stub, which means that incomplete article [11]. We do data collection from Wikipedia, the class distribution of our datasets are shown in Table II.

TABLE II. NUMBER OF ARTICLES BY QUALITY

Class	Total	Vietnamese	Indonesian	Malaysian	Thai	Tagalog
high	1,491	463	479	232	287	30
low	4,142	110	547	925	1,472	1,088

C. *Feature selection*

The complete Wikipedia article should consist of content, infobox, heading, links, image, and citation. Furthermore the number of involved authors and viewer may imply the popularity of that article which may infer the quality of the articles as well. Therefore, we propose to study the feature set as shown in TABLE III

TABLE III. FEATURE SET

No.	Feature name
1	length of article
2	# of infobox
3	# of headings
4	# of wiki links
5	# of external links
6	# of back links
7	# of book citation
8	# of image
9	# of embedded files
10	# of templates
11	# of articles written in other languages
12	# of contributors
13	# of viewers.

We obtain a dataset using Wikipedia API service, which provides access to data and metadata of articles using HTTP, via a URL in a variety formats (including XML, JSON). API service works for every language and is available at the address specified by the template: https://{lang}.wikipedia.org/w/api.php?action={settings}, where {lang} is the language version, {settings} is query settings.

The distribution of features from Thai Wikipedia articles with different quality class are shown in figure 1 to figure 4. The average number of infoboxes is one (see figure 1). The high-quality articles have a lot of wiki links and high number of book citations (see figure 2 and figure 3). The number of

contributors and viewing of visitors may not represent the high quality of the articles (see figure 4).

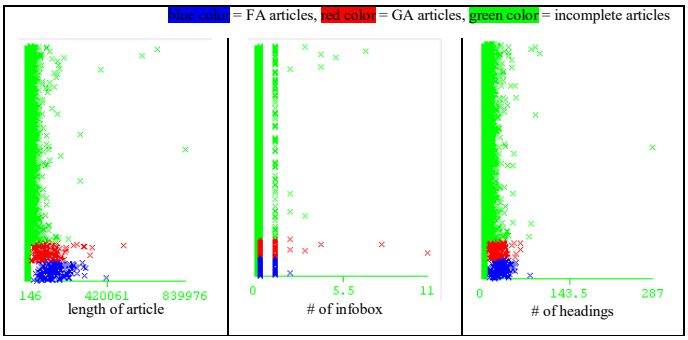

Fig. 1. Distribution of features: length, infobox, and headings in articles with different quality class in Thai Wikipedia.

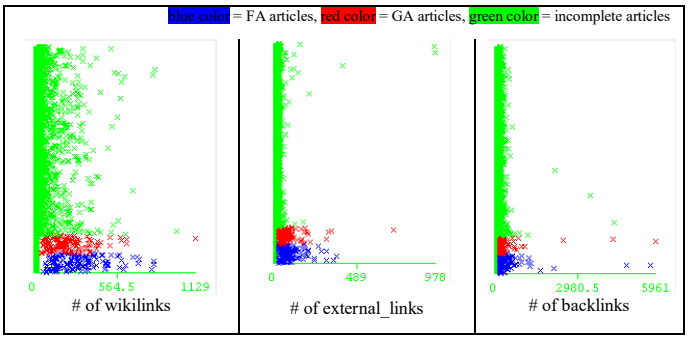

Fig. 2. Distribution of features: wiki links, external links, and back links in articles with different quality class in Thai Wikipedia

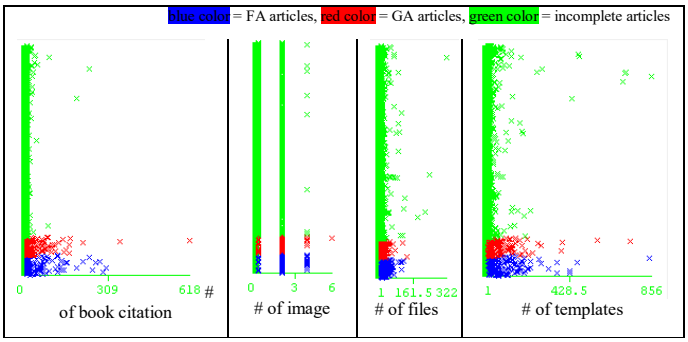

Fig. 3. Distribution of features: book citation, image, files ,and templates in articles with different quality class in Thai Wikipedia

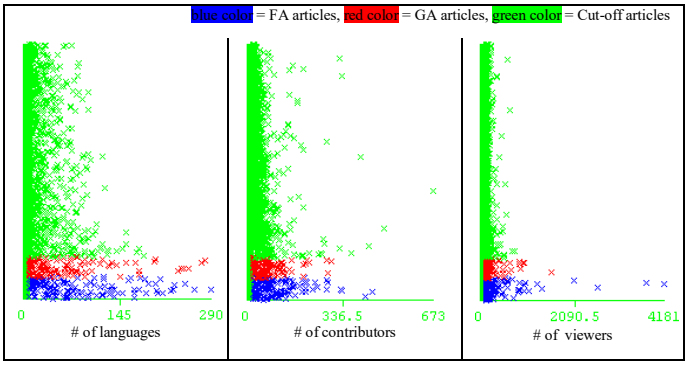

Fig. 4. Distribution of variables: other languages, contributors, and views in articles with different quality class in Thai Wikipedia.

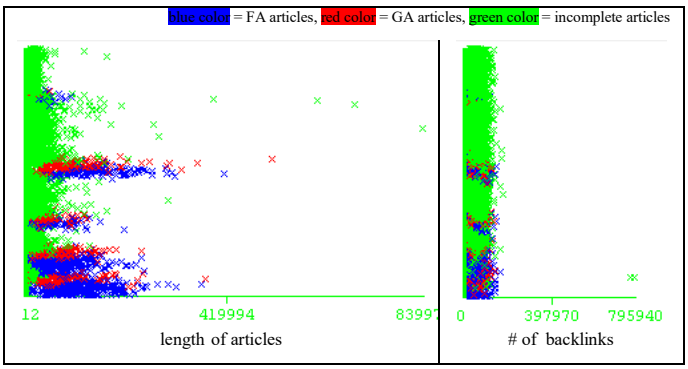

Fig. 5. Distribution of features: length, and back links in articles with different quality class in five studied language articles

We found that the length of the high-quality articles are higher than that of the low-quality articles in every studied language (see figure 5).

IV. Experiment and Evaluation

All experimental results measured in terms of precision, recall and F-measure are shown in Table IV and Table V. We found that Decision tree method outperforms Naïve Bayes. The recall of the high-quality class is 0.873 in mixture model (five language combination dataset), 0.983 in Vietnamese, 0.937 in Indonesian, 0.681 in Malaysian, 0.878 in Thai, and 0.433 in Tagalog (Philippines). For five language combination, the performance of Decision tree is 7% higher than Naïve Bayes. Consider the performance of Naïve Bayes, we found that for the high quality articles Naïve Bayes get less performance than the Decision tree except for Filipino with the recall of 0.433 (NB is 0.6). It provides the lowest performance because the Philippines articles are extremely imbalanced.

The decision trees obtained from the algorithm are shown in figure 6 and 7 for Vietnamese, figure 8 and 9 for Indonesian, figure 10, 11 and 12 for Malaysian, figure 13 and 15 for Thai, figure 14 for Tagalog.

For the decision tree of Thai Wikipedia, we found that if the length of the Thai article is less than 28,805 and # of headings are less than 3 then the article is considered as the low quality article. For the long length article (length > 43095), the model of decision tree reveals that # of backlinks is the key to determine the quality. It is found that when the # of backlinks < 5 then the article has low quality. However the article has high quality if # of external links > 52. This observation is similar to that of Vietnamese and Malaysian (# of external link are 37 and 46, respectively). For Tagalog Wikipedia, if # of external links < 9, the article is considered as the low quality.

For Vietnamese Wikipedia, we found that the low quality of articles can be determined by the length of the article and # of Wikilinks. For the long length article (length > 32751) with high # of heading (more than 11 headings), the article is considered as the low quality article.

For Indonesian Wikipedia, if the length of Indonesian article is less than 14,806, the article is classified as the low quality. But if it is too long (more than 90,500) and # of backlinks are more than 340, the article is classified as the high quality. In case that the length of Indonesian article is less than 11,940, the article is considered as the low quality article.

Vietnamese 1/2

```
length <= 25990
|  nun_languages <= 70
|  |  length <= 8728, length > 8728: high
|  |  |  nun_files <= 2 , // nun_files > 2: low
|  |  |  |  num_templates <= 5, // num_templates > 5: low
|  |  |  |  |  num_wikilinks <= 12: low , // num_wikilinks > 12: high

|  nun_languages > 70
|  |  num_external_links <= 46.5 , // num_external_links > 46.5: high
|  |  |  num_templates <= 127.5: low , // num_templates > 127.5: high
```

Fig. 6. Decision tree (Part 1/2) obtained from Vietnamese Wikipedia

Vietnamese 2/2

```
length > 25990
|  num_backlinks > 712, // num_backlinks <= 712: high
|  |  nun_files <= 22, // nun_files > 22: high
|  |  |  num_templates <= 52.5, // num_templates > 52.5: high
|  |  |  |  nun_languages > 88.5,  // nun_languages <= 88.5: high
|  |  |  |  |  num_contributors > 38.5, // num_contributors <= 38.5: high
|  |  |  |  |  |  num_infobox <= 0.5,  // num_infobox > 0.5: high
|  |  |  |  |  |  |  length <= 87387,  length > 87387: high
|  |  |  |  |  |  |  |  num_external_links <= 66,  /// num_external_links > 66: high
|  |  |  |  |  |  |  |  |  num_citebookweb <= 17.5,  // num_citebookweb > 17.5: high
|  |  |  |  |  |  |  |  |  |  num_views > 33.5,  // num_views <= 33.5: high
|  |  |  |  |  |  |  |  |  |  |  num_image <= 3,  // num_image > 3: high
|  |  |  |  |  |  |  |  |  |  |  |  num_wikilinks <= 634, num_wikilinks > 634: high
|  |  |  |  |  |  |  |  |  |  |  |  |  num_contributors <= 172.5, // num_contributors > 172.5: high
|  |  |  |  |  |  |  |  |  |  |  |  |  |  num_views <= 342, // num_views > 342: high
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  nun_files <= 20.5, // nun_files > 20.5: high
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  num_external_links > 1.5, //num_external_links <= 1.5: high
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  num_headings <= 40.5, // num_headings > 40.5: high
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  length > 32751.5, // length <= 32751.5: high
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  nun_languages <= 229.5, // nun_languages > 229.5: high
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  nun_files > 8.5, // nun_files <= 8.5: high
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  num_wikilinks > 143.5, // num_wikilinks <= 143.5: high
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  num_headings > 11.5: low , //num_headings <= 11.5: high
```

Fig. 7. Decision tree (Part 2/2) obtained from Vietnamese Wikipedia

Indonesian 1/2

```
num_templates <= 17
|  nun_files > 7, // nun_files <= 7: low
|  |  length > 14806, // length <= 14806: low
|  |  |  num_backlinks <= 340
|  |  |  |  nun_files > 21, // nun_files <= 21: high
|  |  |  |  |  num_contributors <= 60.5, // num_contributors > 60.5: high
|  |  |  |  |  |  num_headings <= 26.5, // num_headings > 26.5: high
|  |  |  |  |  |  |  nun_languages <= 37.5 , // nun_languages > 37.5: high
|  |  |  |  |  |  |  |  num_backlinks <= 77.5, // num_backlinks > 77.5: high
|  |  |  |  |  |  |  |  |  length <= 92898.5, // length > 92898.5: high
|  |  |  |  |  |  |  |  |  |  num_external_links <= 19.5, // num_external_links > 19.5: high
|  |  |  |  |  |  |  |  |  |  |  num_citebookweb <= 74.5, // num_citebookweb > 74.5: high
|  |  |  |  |  |  |  |  |  |  |  |  num_wikilinks <= 291.5, // num_wikilinks > 291.5: high
|  |  |  |  |  |  |  |  |  |  |  |  |  nun_files <= 23.5, // nun_files > 23.5: high
|  |  |  |  |  |  |  |  |  |  |  |  |  |  num_views <= 98.5: high , // num_views > 98.5: low

|  |  |  num_backlinks > 340
|  |  |  |  num_external_links <= 41.5, // num_external_links > 41.5: high
|  |  |  |  |  num_citebookweb <= 28, // num_citebookweb > 28: high
|  |  |  |  |  |  nun_files <= 23.5, // nun_files > 23.5: high
|  |  |  |  |  |  |  length <= 90500, // length > 90500: high
|  |  |  |  |  |  |  |  num_wikilinks <= 334, // num_wikilinks > 334: high
|  |  |  |  |  |  |  |  |  nun_languages <= 172.5, // nun_languages > 172.5: high
|  |  |  |  |  |  |  |  |  |  num_image <= 3, // num_image > 3: high
|  |  |  |  |  |  |  |  |  |  |  num_backlinks > 410.5 , // num_backlinks <= 410.5: high
|  |  |  |  |  |  |  |  |  |  |  |  num_views > 74, // num_views <= 74: high
|  |  |  |  |  |  |  |  |  |  |  |  |  num_views <= 478.5: low, // num_views > 478.5: high
```

Fig. 8. Decision tree (Part 1/2) obtained from Indonesian Wikipedia

Indonesian 2/2

```
num_templates > 17

|  num_backlinks <= 9
|  |  num_image <= 0
|  |  |  num_citebookweb > 13, // num_citebookweb <= 13: low (13.0)
|  |  |  |  nun_files <= 13: high, // nun_files > 13: low
|  |  num_image > 0
|  |  |  num_backlinks <= 5: low, // num_backlinks > 5: high
|  num_backlinks > 9
|  |  length <= 15932, // length > 15932: high
|  |  |  num_wikilinks <= 184
|  |  |  |  num_citebookweb <= 2: low
|  |  |  |  num_citebookweb > 2
|  |  |  |  |  num_headings <= 4, // num_headings > 4: high
|  |  |  |  |  |  length <= 9431.5, // length > 9431.5: high
|  |  |  |  |  |  |  num_templates <= 61.5, // num_templates > 61.5: high
|  |  |  |  |  |  |  |  nun_languages <= 7: high, // nun_languages > 7: low
|  |  |  num_wikilinks > 184
|  |  |  |  num_citebookweb <= 95.5, // num_citebookweb > 95.5: high
|  |  |  |  |  num_templates <= 137.5, // num_templates > 137.5: high
|  |  |  |  |  |  num_headings <= 27.5, // num_headings > 27.5: high
|  |  |  |  |  |  |  num_external_links <= 46.5, // num_external_links > 46.5: high
|  |  |  |  |  |  |  |  nun_languages <= 168.5, // nun_languages > 168.5: high
|  |  |  |  |  |  |  |  |  num_wikilinks <= 438.5, // num_wikilinks > 438.5: high
|  |  |  |  |  |  |  |  |  |  num_backlinks <= 317, // num_backlinks > 317: high
|  |  |  |  |  |  |  |  |  |  |  num_image <= 3, // num_image > 3: high
|  |  |  |  |  |  |  |  |  |  |  |  num_contributors <= 95.5, // num_contributors > 95.5: high
|  |  |  |  |  |  |  |  |  |  |  |  |  num_views <= 99.5: low, // num_views > 99.5: high
```

Fig. 9. Decision tree (Part 2/2) obtained from Indonesian Wikipedia

Malaysian 1/3

```
num_templates <= 26
|  num_contributors > 27, //num_contributors <= 27: low
|  |  num_backlinks > 10 , // num_backlinks <= 10:
|  |  |  num_external_links <= 16
|  |  |  |  num_infobox <= 0
|  |  |  |  |  num_citebookweb <= 2, //|  num_citebookweb > 2: high
|  |  |  |  |  |  nun_languages > 54, //nun_languages <= 54: low
|  |  |  |  |  |  |  num_templates <= 10, // num_templates > 10: low
|  |  |  |  |  |  |  |  num_image <= 0
|  |  |  |  |  |  |  |  |  num_headings <= 6: low, // num_headings > 6: high
|  |  |  |  |  |  |  |  num_image > 0
|  |  |  |  |  |  |  |  |  num_wikilinks > 40, //num_wikilinks <= 40: high
|  |  |  |  |  |  |  |  |  |  num_wikilinks <= 125: low, // num_wikilinks > 125: high

|  |  |  num_infobox > 0
|  |  |  |  num_image > 0, // num_image <= 0: low
|  |  |  |  |  num_views > 7, //num_views <= 7: low
|  |  |  |  |  |  num_citebookweb <= 5, // num_citebookweb > 5: high
|  |  |  |  |  |  |  num_templates <= 9: high, num_templates > 9: low

|  |  num_external_links > 16
|  |  |  num_infobox > 0, //num_infobox <= 0: high
|  |  |  |  num_external_links > 31, // num_external_links <= 31: high
|  |  |  |  |  num_external_links <= 103.5, // num_external_links > 103.5: high
|  |  |  |  |  |  length <= 74319: low, // length > 74319: high
```

Fig. 10. Decision tree (Part 1/2) obtained from Malaysian Wikipedia

Malaysian 2/3

```
num_templates > 26
|  num_backlinks <= 9

|  |  num_infobox <= 0
|  |  |  num_contributors <= 3, // num_contributors > 3: low
|  |  |  |  num_contributors > 2, // num_contributors <= 2: low
|  |  |  |  |  num_citebookweb > 17, // num_citebookweb <= 17: low
|  |  |  |  |  |  num_external_links > 19.5, // num_external_links <= 19.5: low
|  |  |  |  |  |  |  num_image > 1, // num_image <= 1: low
|  |  |  |  |  |  |  |  nun_languages > 8, // nun_languages <= 8: low
|  |  |  |  |  |  |  |  |  length > 11940.5, //length <= 11940.5: low
|  |  |  |  |  |  |  |  |  |  num_wikilinks > 52, //num_wikilinks <= 52: low
|  |  |  |  |  |  |  |  |  |  |  num_headings > 6.5, //num_headings <= 6.5: low
|  |  |  |  |  |  |  |  |  |  |  |  nun_files > 2.5, // nun_files <= 2.5: low
|  |  |  |  |  |  |  |  |  |  |  |  |  num_backlinks > 1.5, //num_backlinks <= 1.5: low
|  |  |  |  |  |  |  |  |  |  |  |  |  |  num_wikilinks <= 94.5: high , // num_wikilinks > 94.5

|  |  num_infobox > 0
|  |  |  num_headings <= 14, // num_headings > 14: low
|  |  |  |  num_image > 0, //num_image <= 0: low
|  |  |  |
|  |  |  |  nun_infobox <= 2, // num_infobox > 2:
|  |  |  |  |  num_contributors <= 13, // num_contributors > 13: low
|  |  |  |  |  |  num_templates <= 36, // num_templates > 36: high
|  |  |  |  |  |  |  num_external_links > 31: high , //num_external_links <= 31: low
```

Fig. 11. Decision tree (Part 2/3) obtained from Malaysian Wikipedia

Malaysian 3/3

```
|  num_backlinks > 9
|  |  num_image <= 0
|  |  |  num_infobox > 0, //num_infobox <= 0: low
|  |  |  |  num_external_links <= 37, // num_external_links > 37: low
|  |  |  |  |  num_templates <= 41: low, num_templates > 41: high

|  |  num_image > 0
|  |  |  nun_languages <= 21
|  |  |  |  num_wikilinks > 158, // num_wikilinks <= 158: high (49.0/1.0)
|  |  |  |  |  num_contributors > 5, // num_contributors <= 5: low
|  |  |  |  |  |  num_citebookweb <= 1: low, // num_citebookweb > 1: high

|  |  |  nun_languages > 21
|  |  |  |  num_citebookweb <= 19
|  |  |  |  |  num_views <= 119, // num_views > 119: high
|  |  |  |  |  |  num_infobox <= 0, // num_infobox > 0: low
|  |  |  |  |  |  |  num_headings <= 21: high, // num_headings > 21: low
|  |  |  |  num_citebookweb > 19
|  |  |  |  |  nun_files > 9, // nun_files <= 9: high
|  |  |  |  |  |  num_headings <= 19
|  |  |  |  |  |  |  num_infobox > 0, // num_infobox <= 0: high
|  |  |  |  |  |  |  |  num_contributors <= 20: high, num_contributors > 20: low
|  |  |  |  |  |  num_headings > 19
|  |  |  |  |  |  |  nun_files > 38, // nun_files <= 38: high
|  |  |  |  |  |  |  nun_files > 38
|  |  |  |  |  |  |  |  num_views <= 45, // num_views > 45: high
|  |  |  |  |  |  |  |  |  num_backlinks <= 322.5, // num_backlinks > 322.5: high
|  |  |  |  |  |  |  |  |  |  num_headings <= 31.5: low, // num_headings > 31.5: high
|  |  |  |  |  |  |  |  |  num_backlinks > 322.5: high (0.0|12.0)
```

Fig. 12. Decision tree (Part 3/3) obtained from Malaysian Wikipedia

Thai 1/2

```
num_external_links <= 10

|  num_citebookweb <= 3
|  |  length > 42560, //length <= 42560: low
|  |  |  num_image > 0, //num_image <= 0: low
|  |  |  |  length <= 147730
|  |  |  |  |  num_contributors <= 9, //num_contributors > 9: low
|  |  |  |  |  |  nun_languages > 7.5, //nun_languages <= 7.5: low
|  |  |  |  |  |  |  nun_files > 4.5, //nun_files <= 4.5: low
|  |  |  |  |  |  |  |  num_wikilinks > 49, //num_wikilinks <= 49: low
|  |  |  |  |  |  |  |  |  num_backlinks > 7.5, //num_backlinks <= 7.5: low
|  |  |  |  |  |  |  |  |  |  num_templates > 6.5, //num_templates <= 6.5: low
|  |  |  |  |  |  |  |  |  |  |  num_headings > 5.5, //num_headings <= 5.5: low
|  |  |  |  |  |  |  |  |  |  |  |  num_contributors <= 4.5: low (0.0|37.0)
|  |  |  |  |  |  |  |  |  |  |  num_contributors > 4.5
|  |  |  |  |  |  |  |  |  |  |  |  num_infobox <= 0.5, //num_infobox > 0.5: low
|  |  |  |  |  |  |  |  |  |  |  |  |  num_contributors <= 7.5, //num_contributors > 7.5: low
|  |  |  |  |  |  |  |  |  |  |  |  |  |  num_views <= 1.5, // num_views > 1.5: low
|  |  |  |  |  |  |  |  |  |  |  |  |  |  num_backlinks <= 11.5, //num_backlinks > 11.5: low
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  num_external_links <= 3.5: high, num_external_links > 3.5: low

|  |  |  |  |  nun_files <= 4.5: low, // nun_files > 4.5: high

|  num_citebookweb > 3
|  |  length <= 28805

|  |  |  num_external_links > 9, //num_external_links <= 9: low
|  |  |  |  num_backlinks <= 45.5, //num_backlinks > 45.5: low
|  |  |  |  |  num_views <= 16.5, //num_views > 16.5: low
|  |  |  |  |  |  num_contributors <= 26, //num_contributors > 26: low
|  |  |  |  |  |  |  nun_languages <= 41.5, //nun_languages > 41.5: low
|  |  |  |  |  |  |  |  nun_languages <= 1.5: low
|  |  |  |  |  |  |  |  nun_languages > 1.5
|  |  |  |  |  |  |  |  |  nun_files <= 9.5, nun_files > 9.5: low
|  |  |  |  |  |  |  |  |  |  length > 15867, //length <= 15867: low
|  |  |  |  |  |  |  |  |  |  |  num_headings <= 3.5: low, num_headings > 3.5: high

|  |  length > 28805
|  |  |  num_templates <= 5.5: low (0.0|63.0)
|  |  |  num_templates > 5.5, // num_templates <= 5.5: low
|  |  |  |  num_backlinks > 6.5, //num_backlinks <= 6.5: low
|  |  |  |  num_wikilinks <= 16.5: low, //num_wikilinks > 16.5: high
```

Fig. 13. Decision tree (Part 1/2) obtained from Thai Wikipedia

Tagalog

```
num_headings <= 10: low (1065.0/9.0)
num_headings > 10
|  num_external_links > 9, // num_external_links <= 9: low
|  |  num_contributors > 13, // num_contributors <= 13: low
|  |  |  length <= 92917: high, // length > 92917: low
```

Fig. 14. Decision tree (Part 2/2) obtained from Tagalog Wikipedia

Thai 2/2

```
num_external_links > 10
|  length <= 43095
|  |  nun_languages > 1, // nun_languages <= 1: low
|  |  |  length > 22395, // length <= 22395:
|  |  |  |  num_views > 1, //num_views <= 1: high
|  |  |  |  |  num_backlinks > 15. //num_backlinks <= 15: low
|  |  |  |  |  |  num_contributors <= 60, //num_contributors > 60
|  |  |  |  |  |  |  num_headings > 14, //num_headings <= 14: high
|  |  |  |  |  |  |  |  num_citebookweb <= 30, //num_citebookweb > 30: high
|  |  |  |  |  |  |  |  |  num_external_links <= 52.5, //num_external_links > 52.5: high
|  |  |  |  |  |  |  |  |  |  num_templates <= 98.5, // num_templates > 98.5: high
|  |  |  |  |  |  |  |  |  |  |  num_wikilinks <= 336.5, // num_wikilinks > 336.5: high
|  |  |  |  |  |  |  |  |  |  |  |  num_contributors <= 52.5: low (2.0), //num_contributors > 52.5: high

|  |  |  |  |  |  nun_languages <= 157.5, //nun_languages > 157.5: high
|  |  |  |  |  |  |  num_views <= 508, //num_views > 508: high
|  |  |  |  |  |  |  |  num_backlinks <= 818, //num_backlinks > 818: high
|  |  |  |  |  |  |  |  |  num_external_links <= 59, //num_external_links > 59:
|  |  |  |  |  |  |  |  |  |  num_citebookweb <= 38.5, //num_citebookweb > 38.5: high
|  |  |  |  |  |  |  |  |  |  |  num_headings <= 31.5: low,  // num_headings > 31.5: high

|  length > 43095
|  |  num_backlinks <= 15
|  |  |  num_headings <= 23
|  |  |  |  num_backlinks > 5, // num_backlinks <= 5: low
|  |  |  |  |  num_templates <= 14, //num_templates > 14: high
|  |  |  |  |  num_citebookweb <= 7, //num_citebookweb > 7: high
|  |  |  |  |  |  num_external_links <= 52, //num_external_links > 52: high
|  |  |  |  |  |  |  length <= 65807.5: low, //length > 65807.5: high

|  |  |  num_headings > 23
|  |  |  |  nun_languages <= 70: low, nun_languages > 70: high

|  |  num_backlinks > 15
|  |  |  num_image <= 0, // num_image > 0: high
|  |  |  |  num_headings <= 15
|  |  |  |  |  num_citebookweb <= 26, //num_citebookweb > 26: high
|  |  |  |  |  |  num_external_links <= 44.5, //num_external_links > 44.5: high
|  |  |  |  |  |  |  num_infobox <= 0.5, // num_infobox > 0.5: high
|  |  |  |  |  |  |  |  num_templates <= 63.5: low, // num_templates > 63.5: high

|  |  |  |  num_headings > 15
|  |  |  |  |  num_contributors > 136, // num_contributors <= 136: high
|  |  |  |  |  |  num_views <= 444.5, //num_views > 444.5: high
|  |  |  |  |  |  |  nun_languages <= 28, // nun_languages > 28: high
|  |  |  |  |  |  |  |  num_backlinks <= 818, //num_backlinks > 818: high
|  |  |  |  |  |  |  |  |  num_citebookweb <= 18, // num_citebookweb > 18: high
|  |  |  |  |  |  |  |  |  |  num_wikilinks <= 359.5: high, // num_wikilinks > 359.5: low
```

Fig. 15. Decision tree (Part 2/2) obtained from Thai Wikipedia

TABLE IV. CLASSIFICATION RESULTS PER QUALITY CLASS IN FIVE LANGUAGE VERSIONS USING NAÏVE BAYES..

Class	#Articles	Precision	Recall	F-Measure	Classified	
					high	cleanup
Combination five language versions (Correctly 86.47%, Incorrectly 13.13%)						
high	1,491	0.806	0.645	0.716	961	530
low	4,142	0.881	0.944	0.911	232	3,910
	Avg.	0.861	0.865	0.86		
Vietnam (Correctly 90.57%, Incorrectly 9.42%)						
high	463	0.981	0.901	0.939	417	46
low	110	0.689	0.927	0.791	8	102
	Avg.	0.925	0.906	0.911		
Indonesian (Correctly 83.041%, Incorrectly 16.959%)						
high	479	0.927	0.958	0.943	337	142
low	547	0.962	0.934	0.948	32	515
	Avg.	0.946	0.945	0.945		
Malaysia (Correctly 82.71%, Incorrectly 17.28%)						
high	232	0.6	0.414	0.49	96	136
low	925	0.864	0.931	0.896	64	861
	Avg.	0.811	0.827	0.815		
Thai (Correctly 89.596%, Incorrectly 10.403%)						
high	287	0.706	0.62	0.66	178	109
low	1,472	0.928	0.95	0.939	74	1,398
	Avg.	0.892	0.896	0.893		
Filipino (Correctly 95.34%, Incorrectly 4.65%)						
high	30	0.31	0.6	0.409	18	12
low	1,088	0.989	0.963	0.976	40	1048
	Avg.	0.97	0.953	0.961		

TABLE V. CLASSIFICATION RESULTS PER QUALITY CLASS IN FIVE LANGUAGE VERSIONS USING DECISION TREE.

Class	#Articles	Precision	Recall	F-Measure	Classified	
					high	cleanup
Combination five language versions (Correctly 92.86%, Incorrectly 7.13%)						
high	1,491	0.86	**0.873**	0.866	**1.301**	190
low	4,142	0.954	**0.949**	0.951	212	**3,930**
	Avg.	0.929	0.929	0.929		
Vietnam (Correctly 96.34%, Incorrectly 3.66%)						
high	463	0.972	**0.983**	0.977	**455**	8
low	110	0.924	**0.882**	0.902	13	**97**
	Avg.	0.963	0.963	0.963		
Indonesian (Correctly 93.47%, Incorrectly 6.53%)						
high	479	0.922	**0.939**	0.931	**450**	29
low	547	0.946	**0.931**	0.938	38	**509**
	Avg.	0.935	0.935	0.935		
Malaysia (Correctly 88.24%, Incorrectly 11.75%)						
high	232	0.718	**0.681**	0.699	**158**	74
low	925	0.921	**0.933**	0.927	62	**863**
	Avg.	0.88	0.882	0.881		
Thai (Correctly 95.17%, Incorrectly 4.832%)						
high	287	0.834	**0.878**	0.856	**252**	35
low	1,472	0.976	**0.966**	0.971	50	**1,422**
	Avg.	0.953	0.952	0.952		
Tagalog (Philippines) (Correctly 97.94%, Incorrectly 2.06%)						
high	30	0.684	**0.433**	0.531	**13**	17
low	1,088	0.985	**0.994**	0.989	6	**1,082**
	Avg.	0.976	0.979	0.977		

V. CONCLUSION AND FUTURE WORK

In this paper we have shown that the importance of some article element affects the quality of the information contained in it. In our study we used 13 features of articles and machine learning techniques to come up with a proposal for a quality models. We found that the our proposed feature set play important role in decision tree. We have built the quality models for language edition of Wikipedia and have shown the differences between these models of five languages articles. For future work, we plan to investigate the relevance of the articles in other languages related Thai language. The final goal of our work aims to improve the quality of data in ThaiDBpedia which aims to extract the content from Wikipedia to create the open knowledge for Thai society.

REFERENCES

[1] WikiProject-assessment, https://en.wikipedia.org/wiki/Wikipedia:WikiProject_assessment

[2] S. Javanmardi, "Measuring Content Quality in User Generated Content Systems: a Machine Learning Approach.", Information and Computer Science University of California, Irvine, ProQuest Dissertations, 2011.

[3] Calzada, G.D.l., A. Dekhtyar, "On measuring the quality of Wikipedia articles." In: Proceedings of the 4th workshop on Information credibility, ACM, Raleigh, North Carolina, USA, 2010, pp. 11-18

[4] Blumenstock, J.E.: Size matters: word count as a measure of quality on wikipedia.In:WWW, 2008, pp. 1095–1096.

[5] M. Anderka, "Analyzing and Predicting Quality Flaws in User-generated Content: The Case of Wikipedia", Phd, Bauhaus-Universitaet Weimar Germany, 2013.

[6] Q.V. Dang and C.-L. Ignat, "Quality Assessment of Wikipedia Articles without Feature Engineering," In: 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, ACM, Newark, New Jersey, USA , 2016, pp. 27-30.

[7] G.G. Betancourt, A. Segnine, C. Trabuco, A. Rezgui, and N. Jullien, "Mining team characteristics to predict Wikipedia article quality," In: Proceedings of the 12th International Symposium on Open Collaboration, ACM, Berlin, Germany, 2016, pp. 1-9.

[8] Y. Suzuki and S. Nakamura, "Assessing the Quality of Wikipedia Editors through Crowdsourcing," In: Proceedings of the 25th International Conference Companion on World Wide Web, Canada , 2016, pp. 1001-1006.

[9] K. Saengthongpattana, The Classification of Thai Wikipedia Articles Quality using Concept and Statistical Feature .Ph.D. Thesis, Kasetsart University, 2018.

[10] K. Saengthongpattana, T. Supnithi, and N. Soonthornphisaj," Ontology-Based Classifiers for Wikipedia Article Quality Classification ",In International Symposium on Artificial Intelligence and Natural Language Processing .Rangsit University, Huahin, Thailand, 2017, pp. 21-.29

[11] S. Tamagaw, S. Sakurai., T. Tejima, T. Morita,, N. Izumi, and T. Yamaguchi, "Learning a Large Scale of Ontology from Japanese Wikipedia", In: International Conference on Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM (2010), 2010, pp. 279-286.

[12] Puzzle stub, https://vi.wikipedia.org/wiki/ Tập_tin:Puzzle_stub_ cropped.png