2016 IEEE 2nd International Conference on Collaboration and Internet Computing

Measuring Quality of Collaboratively Edited Documents: the case of Wikipedia

Quang-Vinh Dang
Université de Lorraine, LORIA, F-54506
Inria, F-54600
CNRS, LORIA, F-54506
quang-vinh.dang@inria.fr

Claudia-Lavinia Ignat
Inria, F-54600
Université de Lorraine, LORIA, F-54506
CNRS, LORIA, F-54506
claudia.ignat@inria.fr

***Abstract*—Wikipedia is a great example of large scale collaboration, where people from all over the world together build the largest and maybe the most important human knowledge repository in the history. However, a number of studies showed that the quality of Wikipedia articles is not equally distributed. While many articles are of good quality, many others need to be improved. Assessing the quality of Wikipedia articles is very important for guiding readers towards articles of high quality and suggesting authors and reviewers which articles need to be improved. Due to the huge size of Wikipedia, an effective automatic assessment method to measure Wikipedia articles quality is needed.**

In this paper, we present an automatic assessment method of Wikipedia articles quality by analyzing their content in terms of their format features and readability scores. Our results show improvements both in terms of accuracy and information gain compared with other existing approaches.

I. Introduction

Today Wikipedia is the largest and most common reference source on the Internet. It contains around 40 million articles, among which more than five million articles belong to English Wikipedia. According to Wikimedia Foundation, Wikipedia in general and English Wikipedia in particular get about 14 billion and seven billion page views per month, respectively[1]. A study [1] investigated 1,000 randomly generated search terms in Google and measured the rankings for Wikipedia.org. This study found that Wikipedia is ranked in the first five positions by Google for 96% of queries. This leads to higher probability for Internet users to check the content of Wikipedia [2]. Consequently, it is very important to provide high quality Wikipedia articles.

However, concerns about the quality of Wikipedia have been raised [3], [4]. For instance, the information presented on Wikipedia is not accepted as a reliable source for research by many professors and researchers [5]. The main problem is that, while many articles are of high quality, many others did not receive the desired attention from authors to improve their quality [6], [7].

In order to improve quality of Wikipedia pages, several collaborative projects, such as Collaboration of the Week (CotW), WikiCup and Wikipedia Education Program (WEP), were organized. The success and failure of these projects are discussed in [7]. Wikipedia development team also implemented different kinds of bots[2] to execute several automatic tasks, such as checking if a submitted revision damages a particular Wikipedia page or not [8]. These bots were proved to be efficient in preventing flaws in Wikipedia articles [9]. Nonetheless, measuring the quality of Wikipedia articles is more difficult. A text that does not contain harmful content is valid but might not be of a good quality.

Quality assessment on Wikipedia is being performed by human judgement, based on a small group of experts. In order to assess the quality of a Wikipedia page, several reviewers have to read, review and discuss what quality label should be assigned to this particular page. The process indeed requires a lot of time and effort. Moreover, the process needs to be repeated if the particular page is updated. Currently, the average number of edits per second that are performed on Wikipedia[3] is ten. A manual quality assessment method may not scale well for this editing frequency [10], [11]. An automatic approach for assessing quality is therefore required to support collaboration on Wikipedia. This automatic approach would provide an immediate guidance for readers and search engines to choose high-quality articles, and an immediate feedback for writers and reviewers to have a plan for quality improvement.

In this paper we address the challenge of automatically rating the quality of a Wikipedia articles. We use the following quality class labels defined by Wikipedia ordered from low to high quality: *Stub*, *Start*, *C*, *B*, *GA*, *FA* [10], [7][4]. The description of these quality classes is provided in Table I[5]. Similar to [7], [13], [14] we removed the quality labels *A* and *Bplus* from our analysis as the number of articles belonging to these categories is very small, even less than the number of *FA* articles.

Several research works were proposed for classifying the quality of Wikipedia articles using machine learning algorithms such as *random forest* [14], [13]. These approaches

[1]https://reportcard.wmflabs.org/

[2]https://en.wikipedia.org/wiki/Wikipedia:Bots

[3]https://en.wikipedia.org/wiki/Wikipedia:Statistics

[4]An alternative rating system was proposed by Wikipedia users according to four dimensions, i.e. *complete, trustworthy, well-written, objective*[12], but we do not discuss this rating system in this paper.

[5]https://en.wikipedia.org/wiki/Template:Grading_scheme

978-1-5090-4607-2/16 $31.00 © 2016 IEEE
DOI 10.1109/CIC.2016.42

Class	Description
FA	Professional, outstanding, and thorough; a definitive source for encyclopedic information.
GA	Useful to nearly all readers, with no obvious problems; approaching (but not equalling) the quality of a professional encyclopedia.
B	Readers are not left wanting, although the content may not be complete enough to satisfy a serious student or researcher.
C	Useful to a casual reader, but would not provide a complete picture for even a moderately detailed study.
Start	Provides some meaningful content, but most readers will need more.
Stub	Provides very little meaningful content; may be little more than a dictionary definition. Readers probably see insufficiently developed features of the topic and may not see how the features of the topic are significant.

TABLE I: Description of Wikipedia's quality labels

used each different feature sets for the classification. In this paper, by introducing new nine features to the set presented by [14] and using same machine learning techniques, we present a new classification model to predict the quality of Wikipedia articles. In addition to the features presented in [7] that refer uniquely to the structure of an article, we introduced content-based features of the text such as readability scores. Our model achieves a higher accuracy and information gain compared with other approaches. We showed that our added features play an important role in the performance of the classifier.

The paper is organized as follows. Section II discusses related approaches on measuring quality of Wikipedia articles. Section III presents the set of features of Wikipedia articles that we selected for our model. Section IV presents the data set and the accuracies obtained by the classification algorithms that we used in combination with the proposed feature set. Evaluation of our model using the proposed feature set and the random forest approach are discussed in Section V. Conclusions and directions for future work are presented in Section VI.

II. Related Work

Due to the importance of Wikipedia, many approaches on classifying the quality of Wikipedia articles were proposed. These approaches can be divided into two main categories [14]: using editor-based information such as about article authors, or using article-based information such as about the content or the format of the articles themselves.

A. Using editor-based information

Approaches that used editor-based information analysed information that cannot be computed uniquely from the current content of Wikipedia pages, such as the authors of a particular article, their contributions and the duration of each contribution.

Using the hypothesis that the more reputable an author is, the higher the quality of the articles this author produces, Adler et al. [15] and Javanmardi and Lopes [16] used *reputation* of authors to determine the quality of Wikipedia articles. The result was confirmed in German Wikipedia [17]. The social capital of the editors could also affect the quality of the articles they contributed [18]. Recently, Suzuki applied the idea of using *h-index* on academic ranking for assessing the quality of an article [19].

Another criterion used for assessing the quality of a text is the period of time the text remains stable or is modified by other authors/reviewers. If an article has not been modified significantly for a long time, this article can be considered as mature and of high quality. For instance, Calzada and Dekhtyar [20] used the idea of *stable* article to determine the quality of Wikipedia articles. Wohner and Peters [21] also claimed that a good article should not be modified for a long enough period of time. Biancani [22] showed that there is a strong relationship between the number of words that were not modified for a long period of time and the quality of Wikipedia articles.

Some other research works presented the idea that the quality of Wikipedia articles can be determined based on the interaction between authors and reviewers [11], [10], [13]. For instance, Wilkinson and Huberman [23] showed that a large number of authors and reviewers with an intensive cooperation should lead to high quality articles. Arazy and Nov [24] showed that inequality of editors' local contribution on a particular article, inequality of their global contribution on overall Wikipedia activity levels as well as their coordination affect document quality. Liu and Ram [25] classified editors based on their roles in editing individual Wikipedia articles and identified collaboration patterns among these contributors that are preferable or detrimental for article quality. Li et al. [26] analysed the article-editor network to assess quality of Wikipedia articles. Ruvo and Santone [27] analysed the network of articles in private enterprise wiki systems in order to assess their quality.

B. Using article-based information

The second main approach of assessing quality of Wikipedia articles is to analyse directly the content of Wikipedia articles.

One of the simplest solutions is to measure the length of Wikipedia articles [28]. This solution achieved a very high accuracy in separating between *FA* and *non-FA* articles. Other works considered the writing styles, such as how editors vary the words they used, for assessing articles quality [29], [30].

Dalip et al. [31] analyzed the effect of the feature set comprising text, review and network on the quality of Wikipedia articles. A correlation between this feature set and the quality of Wikipedia articles was performed. Authors claimed that, using the error term of linear regression, the features that describe the structure and style of the articles are the best to distinguish between articles of different quality classes.

Similarly, using content, structure, network and edit history features, Anderka et al. [32] built a binary classifier to predict quality flaws in Wikipedia. They based their approach on the

cleanup tags, which are given by the reviewers who detected the flaws but do not have enough time / expertise to fix them.

Focusing on the feature set that describes the content of Wikipedia articles, Warncke-Wang et al. [14] presented and analyzed a feature set including 17 features. Authors claimed that among these features only a set of 11 should be considered to evaluate the quality of Wikipedia articles. The result is improved in [7].

Based on the work of [7] and [14], Wikimedia Foundation[6] built an online API to predict the quality class of Wikipedia articles called ORES (*Objective Revision Evaluation Service*) [33].

Editor-based approaches are characterised by a high time complexity as they require processing the whole history associated to an article. Moreover, editor-based approaches are indirect predicting methods that rather than considering content information they take into account authors and reviewers related information. For instance, it is not necessary that good authors always write good articles.

We applied the article-based approach, which is faster than the editor-based approach as it uniquely requires processing the current document content. In addition, article-based approaches are direct predicting methods where the quality of a Wikipedia article is determined by its current content, not by its edit history.

We extended the model presented in [7] by adding readability scores to the feature set. Our hypothesis is that the quality of an article depends not only on the structure of an article, but also on how well the article is written. The experiments showed that using readability scores as a part of the feature set can improve the performance of the predicting model. Although some readability scores have been previously studied in [14], [31], [34], they were used in the context of different techniques for measurement of the effect of readability scores on Wikipedia articles quality. Moreover, we propose using some readability scores such as *difficult words* and *Dale-Chall readability score* that have not been investigated by any other study. Later in this paper we show that the proposed readability scores as well as the chosen classification algorithm play a critical role in the performance of the predicting model.

III. Feature selection

In this section we present the features included in our model for assessing the quality of Wikipedia articles. Our hypothesis is that the writing style matters for measuring the articles quality.

We based our model on the one presented in [7]. In addition to the features presented in [7] related to the structure of an article (e.g, does the article has infobox or not, or how many references the article has), we added content-based features of the text. The complete set of features for our model is presented in what follows.

[6]https://wikimediafoundation.org

Variable	Formula
$avg\_sentence\_len$	$\frac{number\_of\_words}{number\_of\_sentences}$
avg_word_len	$\frac{number\_of\_letters}{number\_of\_words}$
$avg\_syllables\_per\_word$	$\frac{number\_of\_syllables}{number\_of\_words}$
$percentage\_of\_difficult\_words$	$\frac{number\_of\_difficult\_words}{number\_of\_words}\%$

TABLE II: Definition of variables used in readability scores

A. Structure-based features

Structure-based features of our model refer to the structure of the document and they are the same as those proposed in [7]. These features are listed below, where the terms inside parentheses represent the variable names used in our model.

- Article length in bytes ($content\_length$)
- Number of references ($num\_references$)
- Number of outlinks to other Wikipedia pages (num_page_links)
- Number of citation templates (num_cite_temp)
- Number of non-citation templates ($num\_non\_cite\_templates$)
- Number of categories linked in the text ($num\_categories$)
- Number of images / length of article (num_images_length)
- Information noise score ($info\_noise\_score$) [35]
- Article has an infobox or not ($has\_infobox$)
- Number of level 2 headings ($num\_lv2\_headings$)
- Number of level 3+ headings ($num\_lv3\_headings$)

B. Content-based features

We added to the model the following content-based features. The variables used in the computation of the content-based features are explained in Table II.

1) Flesch reading score ($flesch\_reading\_ease$): Flesch reading score, or Flesch reading ease [36], is a measure to test how difficult a reading text in English is to understand. Flesch reading ease for a given text is a number between 100 and 0, where higher scores indicate text that is easier to read while lower numbers mark text that is more difficult to read.

$$\begin{aligned} flesch\_reading\_ease = 206.835 \\ - (1.015 \times avg\_sentence\_len) \\ - (84.6 \times avg\_syllables\_per\_word) \end{aligned} \quad (1)$$

2) Flesch-Kincaid grade level ($flesch\_kincaid\_grade$): Flesch-Kincaid grade level [36] for a given English text is a number corresponding to the US grade level required to understand the text. For example, if the score is 9.3, it means that the reader of the text should be ninth grader or higher. Although Flesch reading ease and Flesch-Kincaid grade level use both word length and sentence length as core measures,

they have different weighting factors. These measures are inversely correlated: a text with a high score on the reading ease test should have a low score on the grade-level test.

$$\begin{aligned} flesch\_kincaid\_grade = 11.8 \times avg\_syllables\_per\_word \\ + 0.39 \times avg\_sentence\_len - 15.59 \end{aligned} \tag{2}$$

3) Smog index (smog_index): Smog index [37] of a text estimates the years of education a person needs to understand a given text in English.

$$smog\_index = 3 + \sqrt{polysyllable\_count} \tag{3}$$

The $polysyllable\_count$ is defined as the number of words with more than two syllables.

4) Coleman-Liau index (coleman_liau_index): Coleman-Liau index, or Coleman-Liau readability formula [38] is a linguistic test that measures as Flesch-Kincaid grade the US grade level thought necessary to comprehend a text. As opposed to Flesch-Kincaid grade, Coleman - Liau index relies on characters instead of syllables per word.

$$\begin{aligned} coleman\_liau\_index = 5.88 \times avg\_word\_len - 29.6 \\ \times avg\_sentence\_len - 15.8 \end{aligned} \tag{4}$$

5) Automated readability index (automated_readability_index): Automated readability index (ARI) [39] is another readability score to detect the readability of a given text in English in terms of the US grade level similar to Flesch-Kincaid grade and Coleman - Liau index. ARI and Coleman-Liau index rely on a factor of characters per word, instead of syllables per word as the other listed measures.

$$\begin{aligned} &automated\_readability\_index \\ &= 4.71 \times avg\_word\_len + 0.5 \times avg\_sentence\_len - 21.43 \end{aligned} \tag{5}$$

6) Difficult words (difficult_words): The difficult words score [40] of a given English text is calculated based on how many difficult words appear in a text. A word is considered difficult if it does not appear in a list of 3000 common English words that groups of fourth-grade American students could reliably understand.

7) Dale-Chall score (dale_chall_readability_score): Dale-Chall readability score [41] is another measure for comprehension difficulty when reading a text. This score takes into account the percentage of difficult words in the text as well as the ratio between the number of words and the number of sentences.

$$\begin{aligned} &dale\_chall\_readability\_score \\ &= 0.1579 \times percentage\_of\_difficult\_words \\ &\quad + 0.0496 \times avg\_sentence\_len \end{aligned} \tag{6}$$

8) Linsear write formula (linsear_write_formula): Linsear Write Formula is a readability score initially designed for the United States Air Force to compute the readability of their technical manuals [42]. This score corresponds to the US grade level of a text sample based on sentence length and the number of words used that have three or more syllables.

More precisely, based on a sample of 100 words from the text, where the number of words with two syllables or less is denoted by n_1 and the number of words with three syllables or more by n_2, Linsear Write Formula is calculated as $\frac{n_1+3\times n_2}{number\_of\_sentences\times 2}$ if $\frac{n_1+3\times n_2}{number\_of\_sentences} > 20$ and as $\frac{n_1+3\times n_2}{number\_of\_sentences\times 2} - 1$ in other cases.

9) Gunning-Fog index (gunning_fog): Gunning-Fog index [43] is another readability score to measure the difficulty of a given text in terms of the years of formal education needed to understand the text on a first reading. It is a weighted average of the number of words per sentence, and the number of long words per word.

$$\begin{aligned} gunning\_fog = 0.4 \times (avg\_sentence\_len \\ + percentage\_of\_difficult\_words) \end{aligned} \tag{7}$$

Our proposed model comprises the above readability scores in addition to the 11 features from the original model of [7]. Due to the nature of the above readability scores, we only can apply our model to English Wikipedia.

Several readability scores seem related but this is not a problem for the classification method we chose, i.e. the random forest algorithm, as it can cope with multi-collinearity [44]. Indeed, also other approaches on classification of Wikipedia articles according to their quality used a set of related features. For instance, ORES method used a feature set including the length of the article and the number of images, but also the division of number of images by length, which is derived from the first two features. Furthermore, relationship between features does not necessarily lead to collinearity. For instance, the correlation between $dale\_chall\_readability\_score$ and $difficult\_words$ in our data set is only -0.4.

IV. Predicting models

In this section, we present the data set we used in our experiments and the performances obtained by different classification techniques applied to the set of features presented in the previous section.

A. Data set

We used a set of Wikipedia articles generated through several quality improvement projects ran by Wikipedia as mentioned in Section I. The data set was provided by the authors of [7][7] and it includes the content of 20,489 Wikipedia articles, with corresponding quality labels assigned by Wikipedia reviewers. The distribution of articles within different quality classes is displayed in Table III. As there is no dominating

[7]The data set is available at http://figshare.com/articles/English_Wikipedia_Quality_Asssessment_Dataset/1375406.

Number of *FA* articles	2,415
Number of *GA* articles	3,160
Number of *B* articles	3,209
Number of *C* articles	3,322
Number of *Start* articles	4,110
Number of *Stub* articles	4,273
Total	20,489

TABLE III: Distribution of the data set within different quality classes

quality class in the data set, a naive prediction that predicts every output as the major class, i.e. *Stub*, achieves a low accuracy of 20.8%.

B. Data preparation

The data preprocessing program is written in Python. In order to retrieve the content of a Wikipedia page, we used Wikipedia API[8]. We used the open-source project *wikiclass*[9] to compute structure-based features. The open-source project *textstat*[10] was used to compute content-based features. We collected the content of Wikipedia pages in the data set corresponding to the revision for which the quality labels were assigned.

We applied different classification methods with 5-fold cross-validation techniques to compare the performance of these algorithms on evaluating the quality of Wikipedia articles. 5-fold cross validation is considered as a good practical technique for bias-variance trade-off in evaluating machine learning algorithms [45]. In 5-fold cross validation the entire data set (20,489 articles) is divided into five equal parts (5-fold): four parts are used as a training set and the remaining part as the testing set. This process is repeated five times, each part being used as a testing set alternately.

C. Classification using regression model

In this subsection we present our model using a multiple regression approach [46]. Our dependent variable is the quality class, and the independent variables are the features described above. The target of a multiple regression model is to build a linear function of the dependent variable on other independent variables which best fits to the training data and then use this function to predict the unknown data.

As regression models can be applied only for integer-based values, we converted the quality class to an integer: *Stub* to 0, *Start* to 1, *C* to 2, *B* to 3, *GA* to 4 and *FA* to 6. After using the regression model for predicting the quality class of test data set, we converted back the quality level by rounding.

We achieved an accuracy of 25%, which is not a surprising result as the linear regression is not expected to perform well in classification.

D. Classification using multinomial logistic regression

Multinomial logistic regression [47] is the extended version of logistic regression to classify the data set with more than two possible outputs. Multinomial logistic regression does not require that the dependent variables are continuous.

We achieved an accuracy of 60% on our data set with 5-fold cross-validation.

E. Classification using kNN

k-Nearest Neighbors (kNN) algorithm [48] has been widely used in classification problems. The principle of kNN algorithm is to determine a class of an element (in our case an article) by a majority vote of its neighbors, with the element being assigned to the most common class among its k nearest neighbors. For instance, with k = 3, in order to classify the quality label of an article in the testing data set, first we find 3 nearest articles of this test article in the training data set. If among the 3 nearest neighbors of an article, 2 of them are assigned the quality class *FA* and 1 of them *GA*, the article will be predicted as *FA*. The distance between articles is calculated by using an Euclidean metric in *n-dimension space*, where *n* is the number of features.

In order to apply kNN algorithm for classification of our Wikipedia articles according to their quality, firstly we converted the variable $has\_infobox$ to a numeric value, namely value 0 is assigned to articles without information box, and value 1 to the articles with information box.

The accuracy of cross-validation with kNN is 55%.

F. Classification using CART

In this subsection we present the model using classification and regression tree (CART) [49]. The idea of CART is to build a series of *if - else* decision points to classify the data set, with the goal to minimize the entropy of the training set. A simple example of CART is shown in Figure 1. In this example for instance, to classify the quality label of a particular article, first of all we consider the length of the article: if the $content\_length < 1,000$, we classify this article as Stub, otherwise we consider the variable $num\_references$. If $num\_references \geq 25$, we consider the variable $difficult\_words$, and if $difficult\_words \geq 892$ we classify the article as *FA*, otherwise we classify the article as *GA*.

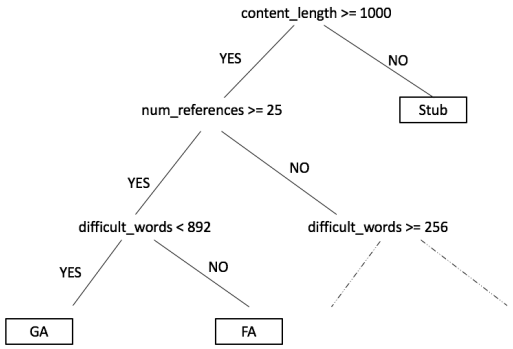

Fig. 1: An example of CART

[8]https://www.mediawiki.org/wiki/API:Main_page
[9]https://github.com/nettrom/Wiki-Class
[10]https://pypi.python.org/pypi/textstat

Using cross - validation with optimized CART, we achieved the accuracy of 48% on our data set.

G. Classification using Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm, which was designed for classification [50]. The idea of Support Vector Machine is to build a hyper plane to classify the data in training set, and then use this hyper plane to predict the quality labels of articles in the testing set.

SVM-based solution achieved an accuracy of 61% on 5-fold cross-validation.

H. Classification using random forest model without readability scores

We tested the performance of the algorithm presented by Warncke et al. [7] based on the source code provided by the authors. In [7], the random forest classification [51] was applied uniquely on the structure-based feature set. Random forest is an improved model of CART where multiple CART are built and used to vote for the output class of a new article. We note that in [7] the content-based features including the readability scores were not used by the random forest model.

By using a 5-fold cross-validation technique on the data set, the algorithm of [7] achieved an accuracy of 58%.

I. Classification using random forest model with the complete set of features

In this subsection, we present the prediction model with random forest with the complete set of features including all structure-based and content-based features. We used a 5-fold cross validation to obtain the optimized parameters for the random forest model, and found that the optimized parameter set is 450 trees with a node size of 4.

This model achieved an accuracy of 64% with 5-fold cross validation on our data set, which is the highest accuracy compared with other existing approaches.

V. Evaluation

In this section we present the performances of our model using the random forest approach. As shown in the previous section, random forest approach provides the best accuracy for our model. We compare performances of our model with other existing approaches by using the following three popular metrics: accuracy, AUC and $NDCG$. We used $NDCG$ score as research approaches that aim ranking articles according to their quality rather than classifying them into quality classes, do not report on accuracy scores.

A. Accuracy & AUC

As our data set is relatively balanced, $accuracy$ is a useful metric to measure classifier performance.

The accuracy score is calculated as:

$$accuracy = \frac{correct\_prediction\_number}{total\_prediction\_number} \tag{8}$$

Approaches concerned with the classification of quality of Wikipedia articles can be divided according to their selected classes for the classification. Some of them distinguish between *FA* and *Start* articles, others between *FA-GA* and *C-Start* articles, and others between all classes.

1) Binary classification: As a fundamental classification, Xu et al [30] classified between two article classes *FA* and *Start*, and they achieved an accuracy of 84%. Lex et al [52] classified the set *FA-GA* classes versus all other classes, and achieved an accuracy of 84%. Wu et al [13] presented two classification results: for the classification between *FA* and *Start* with an accuracy of 85.8% and between two sets of classes *FA-GA* and *C-Start* with an accuracy of 66.4%. For these binary classifications, our method, i.e. random forest with the complete set of features, achieved a very high accuracy: 99.8% on classifying *FA* vs *Start*, 92.7% on classifying *FA-GA* vs *C-Start*, and 91.1% on classifying *FA-GA* vs all other classes.

2) All classes classification: We report the comparison between accuracy and Area Under Curve (AUC) value of different techniques in Table V. The full confusion matrix of our method is displayed in Table IV.

Receiver operating characteristic (ROC) AUC is the measurement of how well does the method behave when the discrimination threshold varies. In order to calculate ROC AUC, firstly the ROC curve is built by plotting true positive rate against false positive rate when the threshold varies and then the area under the curve is computed.

In general, ROC AUC is considered as a more robust measure than accuracy [45]. However, the accuracy metric is important from users' point of view because in applications, a classifier needs to assign one quality label for a particular article without varying the threshold [53]. Therefore, we present both metrics.

AUC is generally defined for binary classification and there does not yet exist a standard way to calculate AUC for multi-class classification [54]. However, the method proposed by Han and Till [55] is widely used. We reported the AUC values calculated by this method in Table V.

Table V showed that our method achieved both higher accuracy and AUC compared to other existing methods. To confirm the improvement in term of statistical significance, we performed McNemar test [56] on our method and the method presented by Warncke et al. [7], which is the second best method in terms of both accuracy and AUC. The McNemar test confirmed that our method is significantly better than the method of Warncke et al. [7], with $p-value < 0.001$. In fact, the difference of 6% on accuracy score between two methods means that we can correct the quality label prediction for about 300,000 English Wikipedia articles.

While random forest was applied for the classification of Wikipedia articles according to their quality, to our knowledge, no prior studies applied some of the previously mentioned machine learning techniques such as multinomial logistic regression and SVM on this purpose. As displayed in Table V, multinomial logistic regression and SVM obtained better accuracy than random forest applied in [7], although the AUC scores are lower. Classification results with good accuracies, but low AUC scores are quite common in the literature [57].

	FA	GA	B	C	Start	Stub	Total	Error Rate
FA	1,816	464	119	12	2	2	2,415	0.2480
GA	640	2,099	167	229	24	1	3,160	0.3358
B	197	505	1,180	771	506	50	3,209	0.6323
C	74	424	580	1,437	757	50	3,322	0.5674
Start	4	56	241	498	2,776	535	4,110	0.3246
Stub	0	1	2	14	549	3,707	4,273	0.1325
Total	2,289	2,961	2,731	3,549	4,614	4,345	20,489	

TABLE IV: Confusion matrix of our method on testing data with cross-validation. Gray cells are correct predictions. Rows are actual quality class. Columns are predicting values of the model. For example, there are 1,816 articles which are predicted correctly as *FA*, and 640 articles which are *GA* and are predicted as *FA*. The last column is the error predicting rate for each class. Because 5-fold cross validation is used, the confusion matrix contains the entire dataset.

Algorithm	RS	Accuracy	$ROC\ AUC$
Linear regression	Yes	25%	0.53
CART	Yes	48%	0.70
kNN	Yes	55%	0.75
Multinomial logistic regression	Yes	60%	0.78
SVM	Yes	61%	0.78
Warncke et al (2015)	No	58%	0.87
Our method	Yes	64%	0.91

TABLE V: Comparison of accuracy and AUC value. RS column indicates whether the corresponding feature set includes readability scores or not.

B. Normalized Discounted Cumulative Gain

In this section, we present the evaluation of our model with the Normalized Discounted Cumulative Gain ($NDCG$) score. $NDCG$ metric proposed by Jarvelin et al. [58] to evaluate ranking systems was used by several studies on the quality assessment of Wikipedia articles [10], [59], [60]. The ranking is done according to the quality of the articles from high to low, i.e. from *FA* to *Stub*. $NDCG$ for the classification of n articles is calculated as:

$$NDCG = \frac{DCG}{iDCG} \tag{9}$$

$$DCG = r_1 + \sum_{i=2}^{n} \frac{r_i}{\log_2(i)} \tag{10}$$

where r_i is the corresponding grade (or gain) of the article which was ranked i^{th} and $iDCG$ (*ideal* DCG), is DCG score for an ideal classification where all the articles are classified correctly. $NDCG$ assigns more importance to the items which were predicted at higher classes such as *FA* than at lower classes such as *Stub*. $NDCG$ score ranges between 0 and 1, with higher values corresponding to a better prediction.

In order to calculate $NDCG$ value for our model, we need to assign a gain score for each article. Using a similar scoring system with [10], we assigned a score of 6 for *FA* articles, 4 for *GA*, 3 for *B*, 2 for *C*, 1 for *Start* and 0 for *Stub*, meaning that *Stub* articles do not contribute to the gain.

Because of the formula of $NDCG$ score, the articles ranked at low level actually have no contribution to the score. For this reason, we usually calculate $NDCG$ score for first k items denoted as $NDCG@k$. For all articles in the test data set, our model achieved a $NDCG$ score of 0.987. The full $NDCG$ score is displayed in Figure 2. Table VI displays the $NDCG$ scores obtained by our model as well as various existing approaches in the literature.

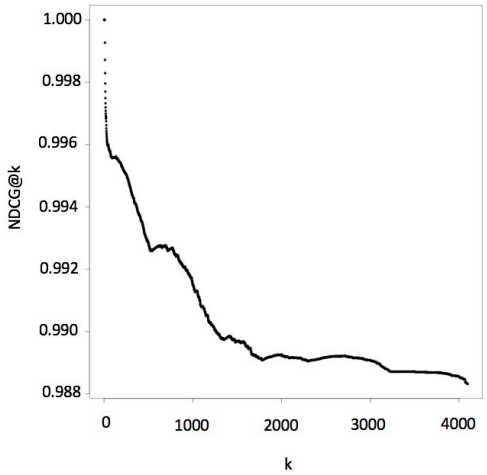

Fig. 2: $NDCG@k$ score of our model

	$NDCG$ score
Hu et al (2007) [10]	0.84
Suzuki (2015) [19]	0.84
Robertie et al (2015) [11]	0.90
Our method	0.987

TABLE VI: $NDCG$ score of different models

The fact that our method achieved a very high $NDCG$ score while the accuracy is not very high is explained by the reason that the $NDCG$ value depends mostly on the correct classification of higher ranked items than lower ranked items. In Table IV, we can observe that our model achieved a low error rate in classifying *FA* articles, and the error rate increases for the classification of lower quality articles, except for *Stub* articles. The $NDCG$ score is calculated mostly based on the classification result of *FA* articles, but does not take into account the high error rate we have in classifying other quality classes, such as *GA*, *B*, *C* and *Start*. For this reason, a model that classifies *FA* articles with a certain level of accuracy

could achieve high $NDCG$ score, despite the fact that it fails classifying other quality labels.

C. Discussion

1) Over-fitting problem: Over-fitting is a critical problem in machine learning. Over-fitting occurs when a model can predict very well on the training set but fails to predict with the testing set.

We used 5-fold cross validation to test the over-fitting problem in our model, as suggested by [45]. As we achieved an accuracy of 64% on both training and testing set we concluded that the over-fitting problem is avoided.

2) Contributions of adding features: We showed that our model improved the accuracy of Wikipedia quality prediction models by adding readability scores. In this section we discuss on how these readability scores contributed to the performance of the model. Three experiments were performed to measure the effect of adding readability scores into the baseline model.

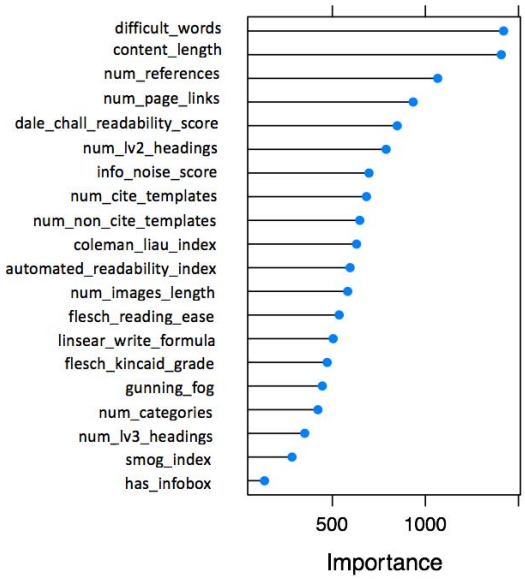

Fig. 3: Feature importance in random forest model

In the first experiment, we measured the feature importance in the model by permuting values of each feature while keeping other features the same, and used a combination of partial least squares and recursive partitioning methods for estimating the contribution of each variable to the model. The importance of each feature in our model for the final performance of the model is displayed in Figure 3.

Among all features of our model, the number of difficult words in the content ($difficult\_words$) had the highest contribution to the accuracy of the model. This is an expected result as difficult words are mostly present in detailed and knowledge intensive articles than in no quality articles. High quality articles require authors to present an in-depth knowledge on the topic. In many cases this leads to the use of a technical

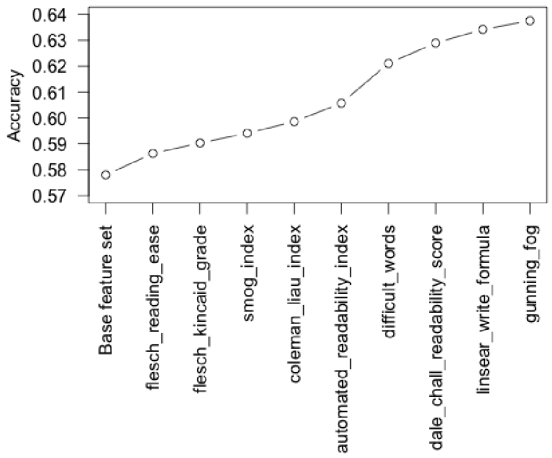

Fig. 4: Accuracy when readability scores are successively added in a cumulative way to the list of features.

language where frequency of occurrence of difficult words is higher.

The variable with the second highest contribution is the length of Wikipedia articles. Even though it is a very simple feature, it plays an important role to determine the quality of the text, as suggested in [28].

The next almost equally important variables are: the number of references ($num\_references$) and the number of links to other Wikipedia pages (num_page_links). Both of these variables show how authors support their content by means of different information sources, which increase the reliability of the article.

The next important variable is Dale-Chall readability score ($dale\_chall\_readability\_score$) that measures how difficult words and sentences are distributed throughout the document. This metric is slightly different than $difficult\_words$. As previously discussed in Section III, in machine learning it is common that using a set of closely related features provides a better accuracy than separately using these features.

In the second experiment, we measured the contribution of added features in a different way: starting from the baseline model, i.e. the model with eleven features [7], readability scores are successively added in a cumulative way to the list of features to study how our model performance changes. The results are presented in Figure 4. The model always performs better when a new readability score is added, although the performance increases are different.

In the third experiment, starting from the baseline model, each readability score was added to the model and then the process was repeated with a different readability score. Each time the experiment was performed, the feature set comprised the baseline structure-based features and one readability score. The results are presented in Figure 5. We observe that adding any readability score improves the performance of the baseline model.

Overall, we can claim that readability scores are important

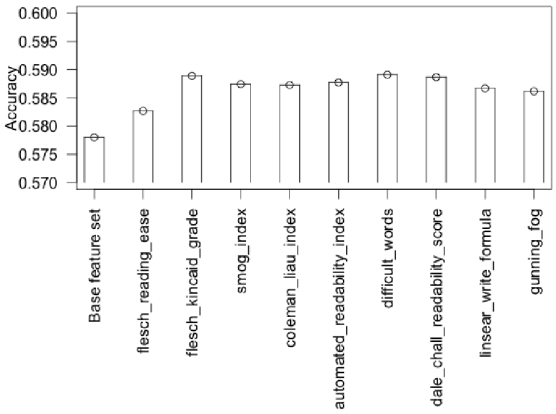

Fig. 5: Accuracy when each readability score is added to the baseline model.

factors to determine the quality of Wikipedia articles and they contributed as much as structure-based scores to the performance of the model.

As we discussed in Section I, the readability scores for measuring quality of Wikipedia articles have been studied by [31], [14], [34]. However, these approaches considered the correlation of readability scores with the quality classes by using linear regression or CART. As we observed in Table V, performances of linear regression and CART on the same feature set and same data set are not as good as random forest. A reason for the difference between performances of linear regression or CART in comparison with random forest is that linear regression and CART fail to deal with collinearity variables [44], [45], [61]. It is important to notice that we used different readability scores than the ones proposed in [31], [14], [34], we investigated the combination of various readability scores and we reported on their importance in the prediction model.

3) Implications for design: In this subsection we present some implications for design for authors to improve the quality of Wikipedia articles and for Wikimedia Foundation on how to provide feedback to authors.

Figure 3 displayed the contribution of each variable to the classification of Wikipedia articles according to their quality classes. However, it is not necessary true that using more difficult words will lead to higher quality articles.

Based on our findings, here are some general suggestions to Wikipedia authors for generating high quality articles:

- Do not hesitate to use technical terms and difficult words if needed.
- Elaborate your ideas: some authors tend to write as concise as possible as they assume that some fundamental knowledge is already known by all other people, which usually is not the case.
- Provide references to support your content.
- Separate the text into small sections rather than using long paragraphs.

Our proposed automatic quality assessment method on Wikipedia articles can serve for several purposes:

- The method achieves very high accuracy on binary classification, so it can notify authors with a high reliability whether their articles belong to a low quality class and they should be improved.
- The method can serve as a measurement for contribution of different authors to a Wikipedia article. For instance, if a Wikipedia article was rated as *Stub* and after the modification of a user the quality of this article becomes *GA*, the contribution of this particular user can be highly rated.

4) Limitation: Our model can be applied uniquely for English Wikipedia articles, because the readability scores we used have been designed particularly for English. In order to extend the model to other languages, we need to adapt the readability scores.

VI. Conclusion

Wikipedia can be considered as one of the most successful user-generated content projects. However, there is a serious concern related to the quality of information in Wikipedia articles. At the time of writing, among more than five million articles of English Wikipedia, only 4,775 (less than 0.1%) articles have been ranked *FA*[11]. If the quality of Wikipedia articles can be automatically computed, we can provide a guidance to readers to select the high quality information and a signal to writers to improve their content.

In this paper, we presented a model to classify the quality class of Wikipedia articles. We showed that in addition to studying the structure-based features by analyzing the content of the articles in terms of their readability, a higher accuracy and information gain can be obtained compared with other approaches. As a future direction of our work we plan to combine manual feature design with automatic feature extraction from deep learning techniques [62] in order to improve classification performance.

VII. Acknowledgements

This research was partially supported by PSPC OpenPaaS::NG project funded by the *Investissements d'Avenir* French government program managed by the *Commissariat général à l'investissement* (CGI).

The authors would like to thank Morten Warncke-Wang, University of Minnesota for sharing the dataset and for the valuable discussions.

References

[1] "Wikipedia: Page one of google uk for 99% of searches," 2012. [Online]. Available: https://www.pi-datametrics.com/wikipedia-page-one-of-google-uk-for-99-of-searches/

[2] "The value of google result positioning," 2015. [Online]. Available: http://chitika.com/google-positioning-value

[3] P. Denning, J. Horning, D. Parnas, and L. Weinstein, "Wikipedia risks," *Communications of the ACM*, vol. 48, no. 12, pp. 152–152, 2005.

[11] https://en.wikipedia.org/wiki/Wikipedia:Featured_articles

[4] T. Simonite, "The decline of wikipedia," *Technology Review*, vol. 116, no. 6, pp. 50–56, 2013.

[5] P. L. Dooley, "Wikipedia and the two-faced professoriate," in *Int. Sym. Wikis*. ACM, 2010.

[6] S. K. Lam and J. Riedl, "Is wikipedia growing a longer tail?" in *GROUP*. ACM, 2009, pp. 105–114.

[7] M. Warncke-Wang, V. R. Ayukaev, B. Hecht, and L. G. Terveen, "The success and failure of quality improvement projects in peer production communities," in *CSCW*. ACM, 2015, pp. 743–756.

[8] J. Segall and R. Greenstadt, "The illiterate editor: metadata-driven revert detection in wikipedia," in *OpenSym*. ACM, 2013, pp. 11:1–11:8.

[9] A. Halfaker, A. Kittur, and J. Riedl, "Don't bite the newbies: how reverts affect the quantity and quality of wikipedia work," in *Int. Sym. Wikis*. ACM, 2011, pp. 163–172.

[10] M. Hu, E. Lim, A. Sun, H. W. Lauw, and B. Vuong, "Measuring article quality in wikipedia: models and evaluation," in *CIKM*, 2007.

[11] B. de La Robertie, Y. Pitarch, and O. Teste, "Measuring article quality in wikipedia using the collaboration network," in *ASONAM*, 2015.

[12] L. Flekova, O. Ferschke, and I. Gurevych, "What makes a good biography?: multidimensional quality analysis based on wikipedia article feedback data," in *WWW*. ACM, 2014, pp. 855–866.

[13] G. Wu, M. Harrigan, and P. Cunningham, "Classifying wikipedia articles using network motif counts and ratios," in *WikiSym*. ACM, 2012, p. 12.

[14] M. Warncke-Wang, D. Cosley, and J. Riedl, "Tell me more: an actionable quality model for wikipedia," in *OpenSym*. ACM, 2013, pp. 8:1–8:10.

[15] B. T. Adler, K. Chatterjee, L. de Alfaro, M. Faella, I. Pye, and V. Raman, "Assigning trust to wikipedia content," in *Int. Sym. Wikis*. ACM, 2008.

[16] S. Javanmardi and C. Lopes, "Statistical measure of quality in wikipedia," in *SOMA*. ACM, 2010, pp. 132–138.

[17] K. Stein and C. Hess, "Does it matter who contributes: a study on featured articles in the german wikipedia," in *Hypertext*, 2007.

[18] K. Nemoto, P. A. Gloor, and R. Laubacher, "Social capital increases efficiency of collaboration among wikipedia editors," in *HT*, 2011.

[19] Y. Suzuki, "Quality assessment of wikipedia articles using h-index," *Journal of Information Processing*, vol. 23, no. 1, pp. 22–30, 2015.

[20] G. D. la Calzada and A. Dekhtyar, "On measuring the quality of wikipedia articles," in *WICOW*. ACM, 2010, pp. 11–18.

[21] T. Wöhner and R. Peters, "Assessing the quality of wikipedia articles with lifecycle based metrics," in *Int. Sym. Wikis*. ACM, 2009.

[22] S. Biancani, "Measuring the quality of edits to wikipedia," in *OpenSym*. ACM, 2014, pp. 33:1–33:3.

[23] D. M. Wilkinson and B. A. Huberman, "Cooperation and quality in wikipedia," in *Int. Sym. Wikis*. ACM, 2007, pp. 157–164.

[24] O. Arazy and O. Nov, "Determinants of wikipedia quality: the roles of global and local contribution inequality," in *CSCW*, 2010, pp. 233–236.

[25] J. Liu and S. Ram, "Who does what: Collaboration patterns in the wikipedia and their impact on article quality," *ACM Trans. Management Inf. Syst.*, vol. 2, no. 2, p. 11, 2011.

[26] X. Li, J. Tang, T. Wang, Z. Luo, and M. de Rijke, "Automatically assessing wikipedia article quality by exploiting article-editor networks," in *ECIR*, ser. LNCS, vol. 9022, 2015, pp. 574–580.

[27] G. D. Ruvo and A. Santone, "Analysing wiki quality using probabilistic model checking," in *WETICE*, 2015, pp. 224–229.

[28] J. E. Blumenstock, "Size matters: word count as a measure of quality on wikipedia," in *WWW*. ACM, 2008, pp. 1095–1096.

[29] N. Lipka and B. Stein, "Identifying featured articles in wikipedia: writing style matters," in *WWW*. ACM, 2010, pp. 1147–1148.

[30] Y. Xu and T. Luo, "Measuring article quality in wikipedia: Lexical clue model," in *SWS*. IEEE, 2011, pp. 141–146.

[31] D. H. Dalip, M. A. Gonçalves, M. Cristo, and P. Calado, "Automatic quality assessment of content created collaboratively by web communities: a case study of wikipedia," in *JCDL*. ACM, 2009, pp. 295–304.

[32] M. Anderka, B. Stein, and N. Lipka, "Predicting quality flaws in user-generated content: the case of wikipedia," in *SIGIR*, 2012, pp. 981–990.

[33] A. Halfaker and D. Taraborelli, "Artificial intelligence service gives wikipedians 'x-ray specs' to see through bad edits," https://blog.wikimedia.org/2015/11/30/artificial-intelligence-x-ray-specs, 2015, accessed: 2016-04-01.

[34] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser, "Assessing information quality of a community-based encyclopedia," in *IQ*, 2005.

[35] ——, "Information quality work organization in wikipedia," *JASIST*, vol. 59, no. 6, pp. 983–1001, 2008.

[36] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom, "Derivation of new readability formulas for navy enlisted personnel," DTIC Document, Tech. Rep., 1975.

[37] G. H. McLaughlin, "Smog grading: A new readability formula," *Journal of reading*, vol. 12, no. 8, pp. 639–646, 1969.

[38] M. Coleman and T. L. Liau, "A computer readability formula designed for machine scoring." *Jour. of Appl. Psychology*, vol. 60, no. 2, 1975.

[39] R. Senter and E. Smith, "Automated readability index," DTIC Document, Tech. Rep., 1967.

[40] J. S. Chall and E. Dale, *Readability revisited: The new Dale-Chall readability formula*. Brookline Books, 1995.

[41] E. Dale and J. S. Chall, "A formula for predicting readability: Instructions," *Educational research bulletin*, pp. 37–54, 1948.

[42] H.-H. Chen, "How to use readability formulas to access and select english reading materials." *Journal of Educational Media & Library Sciences*, vol. 50, no. 2, 2012.

[43] R. Gunning, "The fog index after twenty years," *Journal of Business Communication*, vol. 6, no. 2, pp. 3–13, 1969.

[44] K. Matsuki, V. Kuperman, and J. A. Van Dyke, "The random forests statistical technique: An examination of its value for the study of reading," *Scientific Studies of Reading*, vol. 20, no. 1, pp. 20–33, 2016.

[45] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013.

[46] G. Wilkinson and C. Rogers, "Symbolic description of factorial models for analysis of variance," *Applied Statistics*, pp. 392–399, 1973.

[47] D. Böhning, "Multinomial logistic regression algorithm," *Annals of the Institute of Statistical Mathematics*, vol. 44, no. 1, pp. 197–200, 1992.

[48] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[49] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Wadsworth, 1984.

[50] C. Cortes and V. Vapnik, "Support vector machine," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[51] T. K. Ho, "Random decision forests," in *Document Analysis and Recognition*, vol. 1. IEEE, 1995, pp. 278–282.

[52] E. Lex, M. Voelske, M. Errecalde, E. Ferretti, L. Cagnina, C. Horn, B. Stein, and M. Granitzer, "Measuring the quality of web content using factual information," in *WICOW/AIRWeb*. ACM, 2012, pp. 7–10.

[53] N. Japkowicz and M. Shah, Eds., *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, 2011.

[54] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[55] D. J. Hand and R. J. Till, "A simple generalisation of the area under the ROC curve for multiple class classification problems," *Machine Learning*, vol. 45, no. 2, pp. 171–186, 2001.

[56] M. Eliasziw and A. Donner, "Application of the McNemar test to non-independent matched pair data," *Statistics in medicine*, vol. 10, no. 12, pp. 1981–1991, 1991.

[57] L. A. C. Millard, P. A. Flach, and J. P. T. Higgins, "Rate-constrained ranking and the rate-weighted AUC," in *ECML/PKDD (2)*, 2014.

[58] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, 2002.

[59] X. Qin and P. Cunningham, "Assessing the quality of wikipedia pages using edit longevity and contributor centrality," *AICS*, 2012.

[60] Y. Suzuki and M. Yoshikawa, "Mutual evaluation of editors and texts for assessing quality of wikipedia articles," in *WikiSym*, 2012, p. 18.

[61] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*, 5th ed. John Wiley & Sons, 2012.

[62] Q. V. Dang and C. Ignat, "Quality assessment of wikipedia articles without feature engineering," in *JCDL*. ACM, 2016, pp. 27–30.