# Quality Assessment of Wikipedia Articles without Feature Engineering

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

As Wikipedia became the largest human knowledge repository, quality measurement of its articles received a lot of attention during the last decade. Most research efforts focused on classification of Wikipedia articles quality by using a different feature set. However, so far, no "golden feature set" was proposed. In this paper, we present a novel approach for classifying Wikipedia articles by analysing their content rather than by considering a feature set. Our approach uses recent techniques in natural language processing and deep learning, and achieved a comparable result with the state-of-the-art.

## Keywords

quality assessment, Wikipedia, feature engineering, document representation, deep learning

## 1. INTRODUCTION

Internet has opened the border of traditional libraries: nowadays everyone can participate and contribute to a common human knowledge repository. Wikipedia is a great example of a knowledge resource receiving contribution from a huge number of authors. At the time of writing, there are more than five millions articles in English Wikipedia, and 38 million articles in all languages[1], and the size of English Wikipedia is over 60 times compared with Britannica[2].

However, due to the huge number of contributors[3] and articles, the quality of Wikipedia articles is not equally distributed [19]. Several research works claimed that the quality of centralized human knowledge resources such as books or Britannica are higher than Wikipedia [7, 11].

[1]https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia as on 5 - Jan - 2016.
[2]https://en.wikipedia.org/wiki/Wikipedia:Size_comparisons as on 5 - Jan - 2016.
[3]According to https://en.wikipedia.org/wiki/Wikipedia:Wikipedians, there are more than 100,000 regular Wikipedia editors.



*JCDL '16, June 19-23, 2016, Newark, NJ, USA*



DOI: http://dx.doi.org/10.1145/2910896.2910917

In order to improve the quality of Wikipedia, an effective method is needed for quality assessment of its articles. Wikipedia defines quality classes for its articles, including *FA, A, GA, B, C, Start, Stub* where *FA* is the highest quality class and *Stub* the lowest quality class[4].

Assigning the correct quality class for each Wikipedia article is an important task, as authors and reviewers can be notified to pay more attention for improving the low quality articles, and search engines could promote high quality class articles as query result. However, the high velocity of changes on Wikipedia makes impossible a manual quality assessment of articles by human experts. Therefore, it is important to design an automatic approach for quality assessment of Wikipedia articles.

Existing approaches on this topic [3, 6, 10, 12, 17, 19, 21] are all based on defining a feature set that is believed to describe in the best way the quality of a Wikipedia article. Certain approaches claim that longer articles are of a better quality, some others consider that discussions and interactions among authors and reviewers of an article increase the quality of an article and others consider that the quality of an article is determined by contributions of highly respected authors.

There is no standard rule for selecting features, which is considered as one of the most difficult tasks in machine learning. Moreover, feature selection is language dependent. In this paper, we present a new approach that avoids feature engineering and that determines the quality of an article based on its content. We build a deep neural network model where the input is the full content of the Wikipedia articles, and the output is the quality class of the articles. The same approach can be defined for different language data sets.

We start by presenting related works in quality assessment of Wikipedia articles. We then present our classification model including article representation and the deep neural networks technique that we used for the classification. We then describe the evaluation we performed and we compare our results with state-of-the-art techniques. Finally, we present our concluding remarks and we provide some directions for future work.

## 2. RELATED WORKS

Even though existing research works on automatic quality assessment of Wikipedia articles use a different feature set, they can be classified into two main families: one is analyzing the edit history of an article (for instance, who

[4]The description of each quality class is available at https://en.wikipedia.org/wiki/Template:Grading_scheme.

contributed to the article and the type of their modifications) and the second one is analyzing the article itself (for instance, its length, number of images, presence of an information box).

Belonging to the first family of approaches, [12] measures the quality of Wikipedia articles based on author authority. Using a similar idea, [17] applied authors' *h-index* to measure the quality of articles on Wikipedia. In [18], the authors used both metrics of article's content and authors' authority to measure the quality of Wikipedia articles. However, this research work used a manual evaluation by volunteering students which is not very reliable for verifying the classification. Moreover, the accuracy obtained is not very high. [21] analyzed the edit network around a Wikipedia article to retrieve the information about the quality of that article. [6] presented a model that analysed the collaboration between authors and reviewers on Wikipedia to measure the quality of articles.

On the other hand, as the most simple approach, [3] proposed to use simple word count to evaluate the quality of Wikipedia. Dalip et al [10] analyzed the effect of different feature sets including structure, length, style, review, network and readability in a regression model for measuring the quality of Wikipedia articles and they discussed about a minimal feature set [5]. More recently, [19] used a machine learning model for quality prediction of Wikipedia articles including format-based features such as the number of headings of level 2 of a particular article. Based on the work of [19, 20], Wikimedia Foundation built an online service called ORES to classify the quality class of Wikipedia articles [8], using a set of 24 features for English Wikipedia. This set of features is slightly different for other languages Wikipedia.

Each research work selected and used a different feature set to measure and classify the quality of Wikipedia articles. However, feature selection is mostly based on the heuristic of researchers and so far, there is no "gold - standard" feature set to classify and measure the quality of Wikipedia articles.

In this paper, we claim that the quality of a Wikipedia article should depend on its own content. Certain features can be derived from the article content. Using the full content of Wikipedia articles as the input of training model should avoid missing an important feature that was not manually recognized.

We use the technique *Doc2Vec* [13] to represent Wikipedia articles and a deep neural network to classify their quality. Deep learning is an emerging research field today and, to our knowledge, our work is the first one that applied deep learning for assessing quality of Wikipedia articles. Our approach provides a novel point of view to Wikipedia quality classification.

# 3. CLASSIFICATION MODEL

In this section, we present how to design and feed the content of Wikipedia articles into a neural network.

## 3.1 Article representation

Most machine learning algorithms including neural networks require the input to be represented as a fixed-length feature vector. As Wikipedia articles have different lengths, we need an approach that maps Wikipedia articles to fixed-length feature vectors. The most common fixed-length vector representation for documents is the *bag-of-words* [9] where a document is represented as the bag of its words. However, this approach disregards semantics and even word order.

In this paper, we applied the unsupervised learning algorithm called *Paragraph Vector*, recently known as *Doc2Vec* [13] that learns vector representations for variable-length pieces of texts and overcomes the disadvantages of *bag-of-words* by taking into account the order and semantics of words. In this approach every word and every paragraph are mapped to a unique vector. The paragraph vector is concatenated with several word vectors from the paragraph and trained in order to predict the next word in a text window. While word vectors are shared among paragraphs, paragraph vectors are unique among paragraphs.

We applied the *Doc2Vec* approach where each Wikipedia document corresponds to a paragraph in the above description. While the generated word vectors are not further used, the document vector is given as input for our deep neural network.

## 3.2 Deep neural networks

Deep learning has been successfully applied for several text classification tasks such as Reuters news or sentiment analysis [14].

Neural networks, or artificial neural networks (ANN), are machine learning models inspired by biological neural networks for the estimation of generally unknown functions that depend on a large number of parameters. Neural networks are typically organized in layers made up of a number of interconnected nodes which contain an activation function. Patterns are presented to the network via the input layer, which communicates to one or more hidden layers. The hidden layers perform the actual processing via a system of weighted connections. The hidden layers then transmit the answer to an output layer. A deep neural network (DNN) [2] is defined as an artificial neural network with multiple hidden layers that allows learning abstraction from data.

In our approach, we used a DNN with four hidden layers to learn and classify the representation vectors of Wikipedia articles computed by *Doc2Vec*.

# 4. IMPLEMENTATION AND RESULTS

## 4.1 Implementation

We used the data set contained in around 30,000 English Wikipedia articles which are classified to six quality classes *FA, GA, B, C, Start, Stub* already by Wikipedia reviewers. The data set is provided by Wikimedia Foundations[5]. We separated the data set to training and testing set with the ratio 80/20, similarly to [19] and ORES [8].

We transformed all Wikipedia articles on both training and testing set to *Doc2Vec* vectors by using the library *gensim*[6][7]. The output of the first phase is the collection of vectors for Wikipedia articles in the training and testing set. Therefore, we have a dataset of 30,000 same length vectors. In the second phase, we trained the DNN model on the training set by using *tensorflow*[8], the deep learning library from

[5]The data set is available at http://datasets.wikimedia.org/public-datasets/enwiki/.

[6]http://radimrehurek.com/gensim/

[7]Our hypothesis is that the labeled articles in the training set and the unlabeled articles in the testing set are all completed and available.

[8]https://www.tensorflow.org

Google. Our DNN has four hidden layers, with 2000, 1000, 500 and 200 neurons respectively[9]. The number of neurons is selected as a rule of thumb. The final task is to apply the trained DNN model on the testing vector set, and compare the predicting quality labels with correct values assigned by human judgements.

Currently, no standard methodology exists for constructing an optimal neural network with the right number of layers and number of neurons for each layer. An optimal neural network can be built uniquely empirically [15]. However, randomly choosing a structure for a deep neural network is not a good solution as it leads to performances of a random guess, i.e. a low accuracy of 16.7%.

## 4.2 Results

The predictions obtained by our model are displayed by the confusion matrix in Table 1. The training loss graph in Fig. 1 illustrates the training loss value, i.e. the difference between predicted quality labels and their correct values in the training set, as a function of the number of iterations during the training phase. The graph shows that no local minima is found when the number of training steps reaches a high value (25,000), as the decreasing trend is observed throughout the entire training process.

As the data set we used is balanced, i.e. the number of articles in each class is very close, the *accuracy* metric is suitable to evaluate the classification. Accuracy is defined as a ratio between the number of correct predictions and the total number of articles in the testing data set. The accuracy of our DNN classifier is 55%.

We compare our approach with other popular classification approaches on the same data set. Using the 24 features of ORES as the feature set, k-Nearest Neighbor (k-NN) [1], Classification And Regression Tree (CART) [4] and Random Forest(RF) implemented by ORES, achieved the accuracy of 51%, 48% and 60% respectively. Using the feature set composed of 11 features presented in [19] which is a subset of the 24 features set used in ORES, Random Forest algorithm achieved the accuracy of 58%. The performance of classifiers is summarized in Table 2.

The accuracy of DNN is higher than the one obtained by the k-NN and CART approaches. The lower accuracy of DNN classifier with respect to the RF approach can be explained by the parameter of *Doc2Vec* transformation. Due to our computation power, the size of the vectors was limited to 500, which may lead to the consequence that *Doc2Vec* vectors did not capture all the structure of the Wikipedia articles. Moreover, the low accuracy is also due to our unoptimized DNN, as no standard way exists for constructing a DNN. We can see the improvement from [19] to ORES when more features are added.

To our knowledge, Wikimedia ORES API, which is based on the work of [19, 20] is the only existing approach for classification into all six quality classes. Other works only classified between a subset of classes, such as between *FA* and *Start*[22] with an accuracy of 84%, or between *FA-GA* as a class and the set of *(B, C, Start, Stub)* as another class[16] with an accuracy of 84%, or between *FA-GA* and *C-Start*[21] with an accuracy of 66%. For these binary classification tasks, the DNN approach achieved a very high accuracy compared with previous approaches: 99% to classify between *FA* and *Start*, 86% to classify between *FA-GA* and the other classes and 90% to classify between *FA-GA* and *C-Start*.

We observe that the quality class of a Wikipedia article could be determined by only analyzing its content, so the approach of training the prediction model based on the content and not on feature sets is a promising and interesting approach to be improved in the future. As *Doc2Vec* approach is language independent we expect that our approach can be generally applied to any language Wikipedia.

[9]The implementation is available at https://github.com/vinhqdang/doc2vec_dnn_wikipedia

| | FA | GA | B | C | Start | Stub | Total |
|---|---|---|---|---|---|---|---|
| *FA* | 778 | 148 | 64 | 17 | 6 | 7 | 1020 |
| *GA* | 160 | 554 | 128 | 88 | 23 | 4 | 957 |
| *B* | 87 | 187 | 373 | 237 | 143 | 17 | 1044 |
| *C* | 28 | 112 | 236 | 376 | 181 | 23 | 956 |
| *Start* | 6 | 38 | 119 | 216 | 453 | 133 | 965 |
| *Stub* | 7 | 6 | 20 | 36 | 179 | 701 | 949 |
| Total | 1066 | 1045 | 940 | 970 | 985 | 1006 | 5891 |

Table 1: Confusion matrix of classifying quality classes. Gray cells are correct predictions. Rows (italic) are actual quality class. Columns are predicted values of the model. For instance, there are 778 articles correctly predicted as FA, and 160 articles which are GA and are predicted as FA.

| Classifier | Accuracy |
|---|---|
| CART [4] | 48% |
| kNN [1] | 51% |
| Doc2Vec & DNN | 55% |
| Warncke et al. [19] | 58% |
| Wikimedia ORES [8] | 60% |

Table 2: Accuracy scores of different classifiers on English Wikipedia.

# 5. CONCLUSIONS AND FUTURE WORKS

Feature selection is one of the most difficult task in machine learning. Existing research works proposed different feature sets for measuring quality of Wikipedia articles. Each feature set has its own pros and cons, and there is no "golden feature set". As feature selection process is mostly a manual work, we may never know what feature set is the best for assessing quality of Wikipedia articles.

In this paper, we presented an approach to avoid feature selection process. Our approach follows the process of Wikipedia reviewers: first they read the article and then decide what quality class this article should belong to. Using this approach, no feature selection is required to describe a Wikipedia article. We achieved very high accuracy scores for classification into binary quality classes and an accuracy score comparable with the state-of-the-art Wikimedia ORES service for classification between all quality classes.

As a future work we plan to improve performances of our approach by optimizing the deep neural network's structure.

# 6. ACKNOWLEDGMENTS

Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific inter-

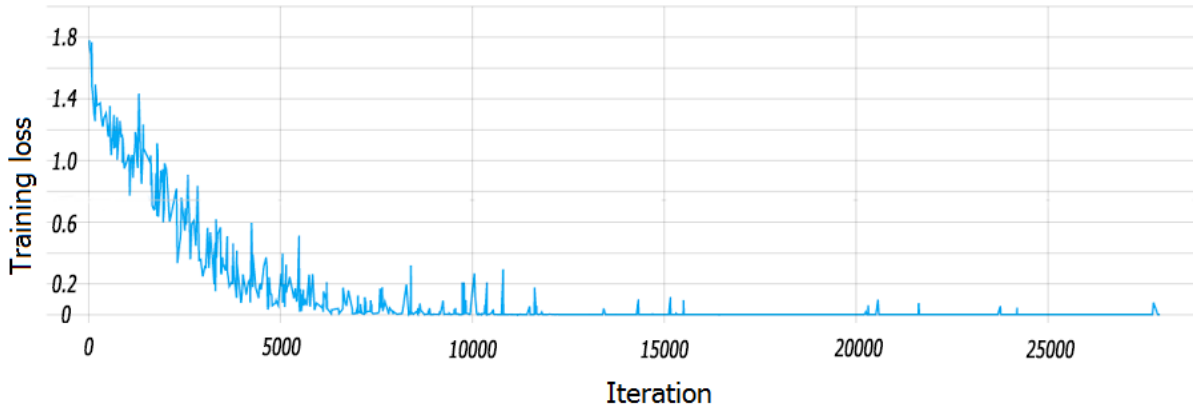


Figure 1: DNN training loss

est group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr).

## 7. REFERENCES


[1] N. S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[2] Y. Bengio. Learning deep architectures for AI. *Found. Trends Mach. Learn.*, 2(1):1–127, Jan. 2009.

[3] J. E. Blumenstock. Size matters: word count as a measure of quality on Wikipedia. In *Proc. of WWW*, pages 1095–1096, 2008.

[4] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and regression trees*. 1984.

[5] D. H. Dalip, H. Lima, M. A. Gonçalves, M. Cristo, and P. Calado. Quality assessment of collaborative content with minimal information. In *Proc. of JCDL*, pages 201–210, 2014.

[6] B. de La Robertie, Y. Pitarch, and O. Teste. Measuring article quality in Wikipedia using the collaboration network. In *Proc. of ASONAM*, pages 464–471, 2015.

[7] P. Dondio, S. Barrett, S. Weber, and J. M. Seigneur. Extracting trust from domain analysis: A case study on the Wikipedia project. In *Proc. of ATC*, pages 362–373, 2006.

[8] 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.

[9] Z. S. Harris. Distributional structure. *Word*, 1954.

[10] D. Hasan Dalip, M. André Gonçalves, M. Cristo, and P. Calado. Automatic quality assessment of content created collaboratively by web communities: a case study of Wikipedia. In *Proc. of JCDL*, pages 295–304, 2009.

[11] L. Holman Rector. Comparison of Wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles. *Reference services review*, 36(1):7–22, 2008.

[12] M. Hu, E.-P. Lim, A. Sun, H. W. Lauw, and B.-Q. Vuong. Measuring article quality in Wikipedia: models and evaluation. In *Proc. of CIKM*, pages 243–252, 2007.

[13] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. In *Proc. of ICML*, pages 1188–1196, 2014.

[14] S. Lee and J. Y. Choeh. Predicting the helpfulness of online reviews using multilayer perceptron neural networks. *Expert Systems with Applications*, 41(6):3041–3046, 2014.

[15] N. D. Lewis. *Build Your Own Neural Network Today*. 2015.

[16] 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 *Proc. of WICOW*, pages 7–10, 2012.

[17] Y. Suzuki. Quality assessment of Wikipedia articles using h-index. *Journal of Information Processing*, 23(1):22–30, 2015.

[18] Y. Suzuki and M. Yoshikawa. Mutual evaluation of editors and texts for assessing quality of Wikipedia articles. In *Proc. of WikiSym*, pages 18:1–18:10, 2012.

[19] 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 *Proc. of CSCW*, pages 743–756, 2015.

[20] M. Warncke-Wang, D. Cosley, and J. Riedl. Tell me more: An actionable quality model for Wikipedia. In *Proc. of OpenSym*, pages 8:1–8:10, 2013.

[21] G. Wu, M. Harrigan, and P. Cunningham. Classifying Wikipedia articles using network motif counts and ratios. In *Proc. of WikiSym*, pages 12:1–12:10, 2012.

[22] Y. Xu and T. Luo. Measuring article quality in Wikipedia: Lexical clue model. In *Proc. of SWS*, pages 141–146, 2011.