Article Quality Classification on Wikipedia: Introducing Document Embeddings and Content Features

Manuel Schmidt, Eva Zangerle
Databases and Information Systems
Department of Computer Science
manuel.schmidt@student.uibk.ac.at,eva.zangerle@uibk.ac.at

ABSTRACT

The quality of articles on the Wikipedia platform is vital for its success. Currently, the assessment of quality is performed manually by the Wikipedia community, where editors classify articles into pre-defined quality classes. However, this approach is hardly scalable and hence, approaches for the automatic classification have been investigated. In this paper, we extend this previous line of research on article quality classification by extending the set of features with novel content and edit features (e.g., document embeddings of articles). We propose a classification approach utilizing gradient boosted trees based on this novel, extended set of features extracted from Wikipedia articles. Based on an established dataset containing Wikipedia articles and quality classes, we show that our approach is able to substantially outperform previous approaches (also including recent deep learning methods). Furthermore, we shed light on the contribution of individual features and show that the proposed features indeed capture the quality of an article well.

CCS CONCEPTS

• **Human-centered computing → Wikis**; *Open source software*; • **Computing methodologies** → *Machine learning*.

KEYWORDS

Wikipedia, Collaborative Information Systems, Information Quality, Classification, Gradient Boosted Trees

ACM Reference Format:
Manuel Schmidt, Eva Zangerle. 2019. Article Quality Classification on Wikipedia: Introducing Document Embeddings and Content Features. In *The 15th International Symposium on Open Collaboration (OpenSym '19), August 20–22, 2019, Skövde, Sweden.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3306446.3340831

1 INTRODUCTION

The Wikipedia platform is one of the largest information sources for people worldwide. Currently, the English Wikipedia features 5.8 million articles which have been shaped by a total of 886 million edits[1]. In December 2018, the English Wikipedia had 7,420 million page views[2], demonstrating how central the Wikipedia has become as a source of information. Consequently, the quality of articles is highly important. Article quality can be considered a combination of multiple factors which range from the trustworthiness of facts (indicated by the number of references and citations), over reading ease to the structure of articles (e.g., [3, 17, 21, 22]). For the different language editions of Wikipedia, the editorial teams behind the editions individually agreed on criteria that allow assessing the quality of an article. For the English Wikipedia, this resulted in a set of quality classes[3] that range from Stub ("A very basic description of the topic.") to Featured Articles ("well-written, comprehensive, well-researched, neutral and stable"). The actual assignment of articles to a specific class is performed manually and in case of changes to an article that may impact the article's quality, the assigned quality class has to be updated manually.

The quality of articles on the Wikipedia platform has been a research topic for more than a decade now (e.g., [3, 18, 19, 21, 22]) and besides characterizing the quality of articles, the automatic classification of article quality has been widely investigated. Approaches for the automatic classification of quality classes can be divided into two main categories: (i) approaches that extract features based on the contents and structure (and possibly, the edit history and information about the editors of the article) of the article to compute a vector representation of the article to subsequently, perform classification based on these vectors (e.g., [6, 22]); and (ii) more recent approaches that rely on Deep Learning approaches to automatically extract meaningful features from the article's raw text to perform the classification task [7, 8]. While the former approaches rely on a wide variety of explicit features like the number of citations or readability indexes, the latter approaches feed the full textual content of the article into a neural network with no explicit feature engineering and rely on the network to compute relevant latent features. This has the advantage that the computation can be performed independently of the language of the article. With explicit feature engineering as performed in the first category, a number of features such as e.g., the number of difficult words contained or the length of the article are dependent on the language of the article. However, a deep learning approach as utilized by Dang and Ignat [8] renders it impossible to attribute the quality of an article to certain features and hence, transparently provide users and editors with information about why an article was attributed to a certain class. This information, in turn, could be highly valuable as it would allow providing editors with hints on how to improve

OpenSym '19, August 20–22, 2019, Skövde, Sweden

ACM ISBN 978-1-4503-6319-8/19/08…$15.00
https://doi.org/10.1145/3306446.3340831

[1] https://en.wikipedia.org/wiki/Wikipedia:Statistics
[2] https://stats.wikimedia.org/EN/SummaryEN.htm
[3] https://en.wikipedia.org/wiki/Wikipedia:Content_assessment

the quality of an article. Such mechanisms could suggest the editor to e.g., add further citations to the article to increase its quality or to work on the readability of the article by lowering the number of complex words—ultimately allowing to provide real-time feedback on article quality and particularly, the various factors that influence article quality and that might be improved on for the current article.

In this work, we propose a classification approach for Wikipedia article quality classes that relies on explicitly defined features. We argue that this allows providing sensible feedback on how the article quality can be improved. However, we propose to not only use an established set of structure and readability features as done in previous work, but we extend the set of features by novel content and edit features (e.g., the document embedding vector of each article computed by doc2vec [13, 14]). As for the classification, we propose to rely on gradient boosted trees (XGBoost) [4], a powerful state-of-the-art tree classification method that also allows computing the relative importance of each feature—ultimately enabling us to derive the most decisive features.

The main contributions of this paper are the following: (i) we present a novel, extended set of features for automated quality assessment of Wikipedia articles that is based on a combination of established features and novel content and edit features to characterize articles and implicitly, their quality; (ii) we show that the novel feature set on the one hand, and gradient boosted trees (XGBoost) as classification algorithm, on the other hand, allows to substantially improve the classification results regarding article classes compared to other state-of-the-art approaches; (iii) we show that the doc2vec vectors are indeed the most important feature in the context of the proposed classification task.

The remainder of this paper is structured as follows. Section 2 discusses related work, while Section 3 introduces the methods and data proposed for classifying the quality assessment classes of Wikipedia. Section 4 presents the experiments conducted to evaluate the proposed approach and discusses the results of these experiments—particularly in comparison to previous approaches. Section 5 concludes the paper and discusses future work.

2 BACKGROUND AND RELATED WORK

The Wikipedia community assesses article quality manually. I.e., the Wikipedia editors categorize articles into seven quality assessment classes [1]. On the English Wikipedia, articles are assigned to these classes based on a set of predefined criteria. We list a short description of each of these classes in Table 1, these descriptions are taken from the Wikipedia article quality schema overview table [1].

Previous research has widely investigated quality measures and classification approaches for Wikipedia articles. Warncke-Wang et al. classify previous research on this matter into editor-based assessment and article-based assessment approaches [22]. As for the approaches focusing on the features of articles, Dalip et al. [5] present a study on a substantial set of textual, structural and editorial features of articles (including length, structure, style and readability, revision history and the social network of editors). Similarly, Blumenstock [3] as well as Wu et al. [24] have shown that the number of words used within an article serves as a good indicator for quality. Warncke-Wang et al [21, 22] proposed a set of features that aim at capturing how well-structured an article is. These features include e.g., the number of headings or the fact whether the article contains an infobox or not. The set of features employed by Warncke-Wang et al. is further extended by Dang and Ignat [6]. Dang and Ignat argue add a set of well-known readability scores and argue that the readability of articles is a crucial feature when it comes to quality. Their evaluations show that the extended feature set is able to outperform the mostly structure-based feature set of Warncke-Wang et al. Dang and Ignat [6] further evaluated different classification techniques based on the feature set of Wang et al. [22]. They assessed regression models, multinomial logistic regression, KNN, classification and regression tree, support vector machine and random forest with random forests providing the best results. Also, Wikimedia's Objective Revision Evaluation Service (ORES) [9] is based on Warncke-Wang et al.'s features and utilizes a gradient boosted tree (XGBoost) approach for the classification of articles as we do in this work.

As for the editor-based assessment of the quality of Wikipedia articles, [11] has shown that effective coordination between editors leads to higher-quality articles on Wikipedia. Similarly, Liu and Ram have shown that coordination between editors and in particular, the specific roles of users ("who does what?") in the process of editing an article influence the quality of articles [12]. Wilkinson and Huberman found that the number of distinct editors highly correlates with the quality of articles [23]. The social interaction of users and editors of Wikipedia has also been examined [2, 3, 12]. While the quality classes employed serve as a good indicator for the quality of articles of a single Wikipedia, Stvilia et al. found that the notion of quality of articles varies significantly among cross-contextual communities (in terms of cultural, social and economic aspects) as formed by the different language editions on Wikipedia [16].

Recently, Dang and Ignat [7] proposed to apply Deep Learning methods for the quality classification task. Hence, they generated a doc2vec representation of each article and fed this representation into a deep neural network to classify article quality. In further work, Dang and Ignat have introduced a Recurrent Neural Network (RNN) approach based on Long-Short-Term-Memory cells which utilizes the words of an article as input for the classification step. They show that this approach is able to outperform the previous approaches utilizing explicit feature engineering in terms of classification accuracy.

In this work, we propose to utilize feature engineering and extend the set of previously utilized features with novel content and two further edit features. We hypothesize that our proposed content features will contribute to the performance of quality classification as these allow to capture the content, semantics and also, elaborateness of the article. We combine this novel set of features used to characterize each article with a gradient boosted trees approach for classification. We reason that firstly, this approach leads to more explainable and transparent results and secondly, the previously employed deep learning approaches have the drawback of substantially longer training and computation times. We argue that even though we rely on a rather simple classification model, we still are able to outperform the deep learning approaches with a substantially lower computation time required.

Class	Description	No. Articles
FA	Professional, outstanding, and thorough; definitive source for encyclopedic information.	4,996
GA	Useful to nearly all readers, with no obvious problems; approaching (but not equalling) the quality of a professional encyclopedia.	5,497
B	Readers are not left wanting, although the content may not be complete enough to satisfy a serious student or researcher.	5,492
C	Useful to a casual reader, but would not provide a complete picture for even a moderately detailed study.	5,492
Start	Provides some meaningful content, but most readers will need more.	5,490
Stub	Provides very little meaningful content; may be little more than a dictionary definition.	5,493

Table 1: Quality classes on Wikipedia (including number of articles of respective category in dataset)

3 METHODS

In the following section, we first present the data underlying our approach. Subsequently, we present the features and classification method proposed for automatically assessing the quality of Wikipedia articles.

3.1 Dataset

For the conducted experiments, we rely on the 2015 dataset provided by Warncke-Wang [20][4], which holds a set of 29,828 English Wikipedia articles. We chose to rely on the 2015 dataset (instead of the 2017 dataset) as it has been widely used in previous work and this allows for directly comparing our results. For each of these articles, the manually assessed quality class is available as ground truth data. Furthermore, pre-defined training- and test data sets are provided. Table 1 presents the quality classes employed on the English Wikipedia and a short description of the criteria that an article has to fulfill to be assigned to the respective quality class (cf. the editing guidelines for the English Wikipedia regarding quality assessments of articles [5]), with FA (Featured Article) and GA (Good Article) describing high-quality articles and Stub describing a very basic article of low quality. We also state the number of articles of the corresponding class label contained in the dataset. Please note that articles labeled as quality class A were removed from the dataset by the original authors [21] as there were too few articles of this class, which would result in a class imbalance. As our approach and experiments are based on the same dataset, we do not consider the A-class either.

For each article contained in this dataset, the page id (identifying the article itself) and the revision id (identifying the version to be considered) are given. Based on this information, we crawl the specified revision of each article and store its contents for performing the proposed quality classification task.

3.2 Features

For the representation of articles, we propose to extend the set of features that have been proposed by previous research. Firstly, we describe the set of established features that we incorporate in our model, before we present the novel features we propose to add.

Along the lines of previous research, we rely on structure and readability features [6, 16, 22]. We list those feature in Table 2 (cf. structure and readability features). Particularly, we rely on the well-established measures developed by Stvilia [16], which have been further extended by Warncke-Wang et al. [22]. Warncke-Wang et al. looked into finding so-called actionable measures, i.e., measures that directly indicate certain flaws within articles to be able to correct these and improve the overall quality of the article. We rely on this list of features for assessing the quality of articles. However, we constrain the set of measures to those directly related to the article (i.e., we exclude the measures Tenure, Completeness, Authority/Reputation, Consistency, and Volatility). This is due to the fact some of these measures require crawling the edit history and all metadata of all editors having contributed to any of the articles in our study. Furthermore, we rely on the readability metrics introduced for Wikipedia article quality classification by Dang and Ignat [6]. Please note that Dang and Ignat extended Warncke-Wang's dataset with readability features and showed that adding readability features improves article classification performance. Here, the intuition is that not only the structure of an article impact its quality, but also how easily readable the article is.

Along the lines of Dang and Ignat [6], we reason that content plays an important role when it comes to judging the quality of articles. Hence, we propose to extend the set of features that have previously been used for article quality classification by content features and additionally, we propose two novel edit features that describe recent edits and their extent.

- **doc2vec:** the intuition behind using the doc2vec [13] embedding vector of the article is that it provides a numeric, latent representation of the document content, its context, and semantics. We hypothesize that adding this comprehensive article representation can be leveraged for getting a better representation of the contents of an article and hence, its quality. After preliminary experiments, we chose to compute a vector representation for each article utilizing 500 latent dimensions.
- **Internal/external links bitmask:** in contrast to previous work, where the sheer number of internal/external links was used, we also aim to characterize the links (pages or base domains, respectively) as a measure to describe the article's content and hypothesize that high-quality articles will refer to similar sources. We hence create a bit vector, where an entry is set to 1 if the article contains a link to the respective page or base domain. To keep the size of the vector manageable, employ a threshold for the minimum number of links to a page/domain to be contained that we determined in preliminary experiments.
- **(Infobox) Categories bitmask:** analogously to the internal/-external links, we also aim to capture the categories employed and hence, again create a bit index for the categories used. We

[4] https://figshare.com/articles/English_Wikipedia_Quality_Asssessment_Dataset/1375406
[5] https://en.wikipedia.org/wiki/Wikipedia:Content_assessment

employ this feature for article as well as infobox categories. The underlying hypothesis for adding this feature is that that articles within a common quality class also share the same categories in the article.

- **Templates bitmask:** along the same lines as the categories employed, we also capture the templates used for structuring the article. Here, the underlying assumption is that templates are used to structure the article and to mark certain flaws in articles and hence, looking into which templates are featured in an article, should provide us with additional content information.
- **POS tags count:** As a further content feature, we propose to use the count of POS tags. The intuition here is that this should give us an indication of how an article is written from a stylometry point-of-view [10] (e.g., we can capture how many adverbs or adjectives are contained).
- **Length of sections:** sections have been shown to be a good indicator for the structure of an article [22], while at the same time, article quality has also been correlated to article length [3]. We propose to combine these findings and to also use the length of sections to describe an article in a vector representation.
- **Timestamps edit history:** in previous work, the currency of an article has often been measured by the time since the last edit [22]. We propose to extend this by not only using the timestamp of the last edit but of the last 100 edits. This allows getting a more comprehensive picture on the recent edit frequency of the article.
- **Diffs edit history:** While the last edits' timestamps allow measuring the article's currency, they do not provide information about the extent of the changes performed. Hence, we propose to utilize the vector differences between the tf/idf vectors of the last 100 versions of the article.

For the computation of the information noise score, we made use of the NLTK Porter Stemmer[6] and also make use of NLTK's set of stopwords[7]. As for the readability-based features, we rely on the features proposed by Dang and Ignat [6]. For the extraction of those features, we rely on the Python TextStat library[8] and for the computation of the doc2vec vectors, we relied on the gensim library[9].

After having computed the individual features, we concatenate the features into a feature vector that describes the given article. These vectors serve as input for the article quality class classification approach.

3.3 Classification

Based on the article features described in the previous section, we aim to assign each article (rather, the computed vector representation of the article) a quality class. Along the lines of the ORES system [9], we propose to utilize gradient boosted trees for this task and particularly, rely on the XGBoost approach [4]. In principle, gradient boost trees are a tree classifier that relies on boosting to combine a set of weaker tree models to a more comprehensive and accurate model. This is done by analyzing the features, their importance and performance of the weaker trees to iteratively reduce misclassifications of the previous model to improve the overall performance.

Proposed Features
doc2vec representation
Internal links bitmask
External links bitmask
Categories bitmask
Templates bitmask
Infobox categories bitmask
POS tags counts
Length of sections
Timestamps of last 100 edits
tf/idf differences of last 100 edits
Structural Features [21, 22]
Article length in bytes (log-transformed)
Number of references (log-transformed)
Number of links to other articles (log-transformed)
Number of citation templates
Number of non-citation templates (log-transformed)
Number of categories linked in the text
Number of images / length of article
Information noise score
Article has infobox or not
Number of level 2 headings
Number of level 3+ headings
Number of sections (total)
Number of citations (log-transformed)
Readability Features [6]
Flesch reading score
Flesch-Kincaid grade level
Smog index
Coleman-Liau index
Automated readability index
Difficult words
Dale-Chall score
Linsear write formula
Gunning-Fog index

Table 2: Overview of features utilized for Wikipedia article quality classification

4 EXPERIMENTS AND RESULTS

In the following, we present the experiments conducted to evaluate the performance of the proposed quality prediction approach.

4.1 Experimental Setup

We base our experiments on the set of articles presented in Section 3.1 as these have been used in previous work as well. The dataset provided by Warncke-Wang et al. [20] contains articles described by the revision ID where the article first belonged to a given quality class and the quality class. For each of these articles, we

[6] http://www.nltk.org/api/nltk.stem.html
[7] http://www.nltk.org/nltk_data/
[8] https://github.com/shivam5992/textstat
[9] https://radimrehurek.com/gensim/models/doc2vec.html

fetch the content of the given revision using the MediaWiki-API[10]. As some of the given revisions have been deleted, we fetch the first available more recent revision for the article (as already proposed by Warncke-Wang et al. [21]). This provides us with a dataset of 29,366 articles. The distribution of the manually assessed quality of these articles is shown in Table 1. Based on the crawled data, we extract the features presented in Section 3.2. Based on the resulting vector representations, we perform the classification.

For the evaluation, we perform a 5-fold cross-evaluation for the proposed approach along the lines of previous research [6, 8] based on Warncke-Wang et al.'s dataset as described in Section 3.1. Hence, we randomly split the dataset into five folds and repeat the evaluation five times, with each fold serving as the test dataset once. For tuning the gradient boosted trees approach, we perform a grid search to find the best parameters for XGBoost.

As for the metrics used for evaluating the proposed approaches, along the lines of previous research [6, 8] we rely on the accuracy metric.

4.2 Evaluated Methods

For evaluating and contextualizing the proposed approach, we compare our proposed approach to the following state-of-the-art approaches. Furthermore, we add a deep learning-based baseline that allows assessing the performance of XGBoost given the same feature set. Please note that we refer to each approach by a short name and add the utilized features (or rather, input) in parenthesis, before we shortly describe the approach.

- **XGBoost (all features):** Proposed approach utilizing the features stated in Section 3.2 and using gradient boosted trees (implemented by the XGBoost library) for quality classification.
- **RNN-LSTM (article text):** Dang and Ignat's [8] approach uses a recurrent neural network architecture based on long short term memory units to classify article quality and uses the set of tokens contained in the article as input. It represents the currently best performing approach towards article quality classification.
- **Random Forest (structural, readability features):** Random Forest classification based on Dang and Ignat's [6] feature set including structural and readability features.
- **DNN (doc2vec vectors)**: Dang and Ignat's [7] approach for article quality classification utilizes doc2vec vector representations as input for a classification approach based on deep neural networks.
- **DNN (all features):** Along the lines of the DNN doc2vec approach, we utilize a deep neural network featuring five fully connected layers with decreasing number of outputs to compute the classification task (ReLu activation, dropout for each layer, Adam optimizer, batch size of 64).
- **ORES (Warncke-Wang features):** Wikimedia's Objective Revision Evaluation Service (ORES) also provides a quality classification service[11]. This service is based on Gradient boosted trees (XGBoost) and based on the work and features by Warncke-Wang et al. [21, 22]. We base the results presented here on the information provided on the ORES evaluation website[12].

Please note that for the baseline methods RNN-LSTM, Random Forest and DNN Doc2Vec, we report the accuracy values as reported in the original papers as re-running the proposed deep learning experiments would have exceeded our computational capabilities. However, we argue that our experiments were conducted on the same dataset and hence, we consider it reasonable to report the original results here.

4.3 Results and Discussion

In the following section, we first present the results of the conducted classification experiments and subsequently, investigate the relative importance of the utilized features.

Approach	Accuracy
XGBoost all features	73%
RNN-LSTM	68%
DNN all features	67%
Random Forest	64%
ORES	62%
DNN Doc2Vec	55%

Table 3: Classification accuracy results for all methods (sorted by accuracy)

4.3.1 Classification Performance. Table 3 depicts the results of the evaluation of the quality classification experiments conducted. As can be seen, the best performing approach is the proposed XGBoost approach utilizing the full novel feature set, reaching a classification accuracy of 73%. In comparison, the second-best approach is the Recurrent Neural Network based on LSTMs, achieving an accuracy of 68%. The third approach is the neural network for classification based on the full feature set proposed, reaching an accuracy value of 67%, which further underscores the performance of the proposed features. Dang and Ignat's Random Forest- based approach utilizing structure and readability features reaches an accuracy of 64%, while ORES (notably, utilizing XGBoost) achieves 62% accuracy. Furthermore, the neural network-based approach utilizing solely doc2vec feature vectors achieves 55% accuracy. When comparing the results of the proposed approach to the accuracy values of the ORES approach, which utilizes XGBoost for classification as well, we can observe that our extended feature set is indeed better able to capture quality aspects of articles. Similarly, this holds for the random forest approach employed by Dang and Ignat based on structural and readability features.

In Table 4, we depict the confusion matrix of a quality classification performed by the proposed approach. For the creation of this matrix, we again rely on a five-fold cross evaluation and subsequently compare the resulting predicted classes with the actual ground truth classes. As can be seen, the proposed approach is

[10] http://www.mediawiki.org/wiki/API:Main
[11] https://www.mediawiki.org/wiki/ORES#Article_quality
[12] https://meta.wikimedia.org/wiki/Objective_Revision_Evaluation_Service/wp10

Class	FA	GA	B	C	Start	Stub
FA	**4,267**	664	24	1	0	0
GA	520	**3,222**	861	338	34	0
B	41	754	**3,398**	444	347	7
C	41	489	1,249	**2,606**	569	35
Start	1	17	251	347	**3,994**	390
Stub	2	4	3	45	220	**4,181**

Table 4: Confusion matrix for XGBoost approach on all proposed features.

able to classify the majority of all articles correctly (the number of wrongly classified articles is substantially lower than the number of correctly classified articles). We also observe that the classes on the upper and lower end of the quality spectrum (FA, Stub) achieved the best results with the lowest number of misclassifications. In fact, for the Featured Article class, only 689 out of the 4,956 articles are not classified correctly (13.90% of all FA articles). Similarly, 6.15% of all Stub articles are wrongly classified. However, we also observe that our approach performs worse when having to classify the mid-quality feature classes such as B or C. For instance, the C class achieves the worst results with only 47.77% of all articles of the B class correctly classified. From the confusion matrix we see that the wrongly classified C-class articles are mostly assigned to the neighboring classes B and Start. Similar behavior has also been observed by previous studies [6, 22]. This seems quite natural as the boundaries between those classes are rather subtle and the according criteria are formulated in a rather shallow form. For instance, the criteria proposed by the Wikipedia community for English C class articles are the following: "The article cites more than one reliable source and is better developed in style, structure, and quality than Start-Class, but it fails one or more of the criteria for B-Class. It may have some gaps or missing elements; need editing for clarity, balance, or flow; or contain policy violations, such as bias or original research. Articles on fictional topics are likely to be marked as C-Class if they are written from an in-universe perspective. It is most likely that C-Class articles have a reasonable encyclopedic style." This formulation leaves room for subjective assessment of the community. We argue that consequently, these classes are also hard to assess for human users. This in turn also shows in the results of our evaluations. The ground truth data utilized for this evaluation is a manual classification of each article based on the above mentioned rather vague criteria. As our tree classifier is trained on this data, we naturally observe the continuation of the lack of clear criteria and hence, blurring boundaries between these mid-quality classes in our automatic classification approach. Hence, we consider these misclassifications into neighboring classes of lower importance than getting the overall picture of article classifications right.

To conclude, our experiments have shown that the proposed, novel features are indeed able to capture the quality of articles well and that they are able to outperform the state-of-the-art approaches in the field of article quality classification.

4.3.2 Feature Importance. In the previous experiment, we have shown that our proposed set of features in combination with the XGBoost classifier is able to outperform other state-of-the-art approaches. To complement these findings with an evaluation of the impact of the proposed features on the classification performance, we discuss feature importance of these features in the following. Therefore, we utilize the gain of each feature in the XGBoost model [4], which is a measure for the improvement in accuracy when adding a split on the given feature to the tree. This gain is computed for each feature in every tree of the trained model and is then averaged to a final gain value for each feature. Figure 1 shows the information gain for the top-10 features. Inspecting the top-10 features impacting article quality, nine of the newly proposed features are included.

We observe that the most important feature is the doc2vec feature vector, contributing a substantially higher information gain than the other features. These findings suggest that the doc2vec vectors is indeed able to capture the quality of an article by incorporating and representing content, context, and semantics of the article as computed by the embedding approach. Notably, the other features provide a substantially lower information gain. Furthermore, we also observe that the features describing the employed templates also contribute to the quality classification. We lead this back to two factors: first, templates contribute to the structure of articles and secondly, as already noted by [22], templates are also used by the community to mark specific article flaws, which naturally contributes to the classification performance. Furthermore, our proposed extended features describing the recency and the extent of recent changes (Diffs edit history and Timestamps edit history) also are among the top features. The same holds for POS tags, section length, infobox categories, and internal links. We consider these features as representative for the elaborateness (e.g., length of sections) and the extent to which an article is embedded in the Wikipedia environment (e.g., internal links, infobox categories). Dang and Ignat also looked into feature importance and found that for their model, the number of difficult words, content length and the number of references are the most important features in a random forest classification. In our model, we also find the number of references among the top-10 features.

As for the practical implications of this work, we believe that it can serve as a good indicator for article quality. For example, such a classification approach could inform a recommender system that supports Wikipedia editors in assessing an article's quality by providing a recommended quality class. Another possible application scenario could be a recommender system that provides editors with information about which features of the currently edited article need to be improved to reach a higher quality (e.g., the number of citations or the structure of the article).

Generally, this evaluation confirms our hypothesis that the proposed feature and hence, the content of an article, play an important role when it comes to judging article quality.

4.3.3 Limitations. We acknowledge the fact that the proposed approach requires determining and fixing the set of features utilized for the classification step. This is in contrast to deep learning-based methods where the full text of the article to be classified can directly be used as input. This need for computing the set of features

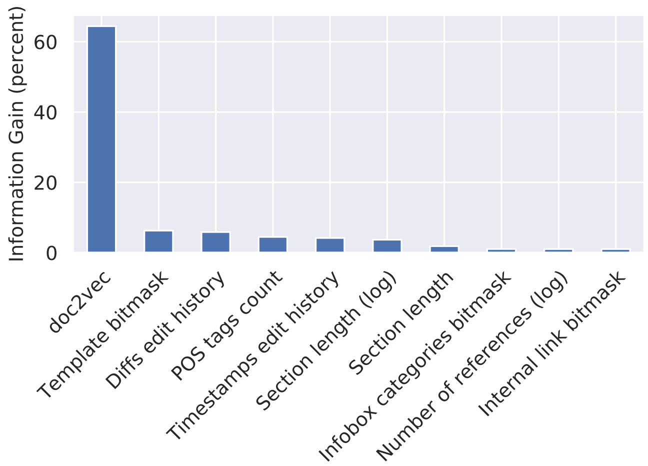

Figure 1: Feature importance (information gain) for top-10 features

to describe the article and its content in advance naturally has a number of drawbacks. Firstly, the set of features to be used has to be fixed before the classification step and does not leave room for computing latent features characterizing an article. Secondly, some of the features we utilized are language-dependent such as e.g., the length (e.g., word vs. syllable languages) of an article or the number of difficult words, which requires a language-specific list of difficult words. This requires a language-specific implementation of these features and possibly, an adaptation of the approach for each language.

5 CONCLUSION

We presented a novel approach for article quality classification on Wikipedia that makes use of an extended set of features utilized to describe an article's quality. Particularly, we propose to make use of the doc2vec vector representation of the article. Furthermore, our approach relies on gradient boost trees, a special form of random forest classifiers. Our experiments showed that including the doc2vec representation of an article to describe its content has a high impact on the classification performance. By relying on gradient boosted trees, we rely on a highly performant and transparent classification model. The experiments conducted show that the combination of the proposed novel feature set and XGBoost classification is able to outperform current state-of-the-art (deep learning) approaches for this task by 5%. We argue that the proposed approach not only provides us with more accurate classification results, but also with a more transparent classification procedure that allows using feature importance to e.g., provide feedback on possible improvements regarding article quality. Furthermore, XGBoost is a highly scalable classification approach, allowing to compute classifications with substantially less resource-intensive than deep learning approaches (with the exception of computing the doc2vec feature vectors, which nevertheless can be computed efficiently).

Future work will include further refining the proposed features and particularly, looking into stylometric features that allow to describe the writing style of an article. In natural language processing, there are numerous features that aim at describing the writing style of authors aiming at performing tasks like authorship attribution [15]. In future work, we aim to borrow from this stream of research and look into to which extent such lexical, syntactic or semantic features may also be used for determining the quality of Wikipedia articles. Furthermore, we also aim to look into different variations of the computed word embeddings.

REFERENCES

[1] 2017. Wikipedia: Template: Grading scheme. (2017). https://en.wikipedia.org/wiki/Template:Grading_scheme

[2] Ofer Arazy, Oded Nov, Raymond Patterson, and Lisa Yeo. 2011. Information Quality in Wikipedia: The Effects of Group Composition and Task Conflict. *J. Manag. Inf. Syst.* 27, 4 (apr 2011), 71–98. DOI: http://dx.doi.org/10.2753/MIS0742-1222270403

[3] Joshua E Blumenstock. 2008. Size Matters: Word Count as a Measure of Quality on Wikipedia. In *Proceedings of the 17th International World Wide Web Conference.* ACM, 1095–1096. DOI: http://dx.doi.org/10.1145/1367497.1367673

[4] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16).* ACM, New York, NY, USA, 785–794. DOI: http://dx.doi.org/10.1145/2939672.2939785

[5] Daniel Hasan Dalip, Marcos André Gonçalves, Marco Cristo, and Pável Calado. 2009. Automatic Quality Assessment of Content Created Collaboratively by Web Communities: a Case Study of Wikipedia. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries.* 295–304. DOI: http://dx.doi.org/10.1145/1555400.1555449

[6] Quang-Vinh Dang and Claudia-Lavinia Ignat. 2016. Measuring Quality of Collaboratively Edited Documents: the case of Wikipedia. In *Proceedings of the 2nd IEEE International Conference on Collaboration and Internet Computing (CIC-16).* Pittsburgh, United States. https://hal.archives-ouvertes.fr/hal-01388614

[7] Quang Vinh Dang and Claudia-Lavinia Ignat. 2016. Quality assessment of wikipedia articles without feature engineering. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries.* ACM, 27–30.

[8] Quang-Vinh Dang and Claudia-Lavinia Ignat. 2017. An End-to-end Learning Solution for Assessing the Quality of Wikipedia Articles. In *Proceedings of the 13th International Symposium on Open Collaboration (OpenSym '17).* ACM, New York, NY, USA, Article 4, 10 pages. DOI: http://dx.doi.org/10.1145/3125433.3125448

[9] Aaron Halfaker and Dario Taraborelli. 2015. Artificial intelligence service "ORES" gives Wikipedians X-ray specs to see through bad edits. (2015). https://blog.wikimedia.org/2015/11/30/artificial-intelligence-x-ray-specs/

[10] David I Holmes. 1998. The evolution of stylometry in humanities scholarship. *Literary and linguistic computing* 13, 3 (1998), 111–117.

[11] Aniket Kittur and Robert E. Kraut. 2008. Harnessing the Wisdom of Crowds in Wikipedia: Quality Through Coordination. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work (CSCW '08).* ACM, New York, NY, USA, 37–46. DOI: http://dx.doi.org/10.1145/1460563.1460572

[12] Jun Liu and Sudha Ram. 2011. Who does what: Collaboration patterns in the wikipedia and their impact on article quality. *ACM Transactions on Management Information Systems (TMIS)* 2, 2 (2011), 11.

[13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[14] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems.* 3111–3119.

[15] Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology* 60, 3 (2009), 538–556. DOI: http://dx.doi.org/10.1002/asi.21001 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.21001

[16] Besiki Stvilia, Abdullah Al-Faraj, and Yong Jeong Yi. 2009. Issues of cross-contextual information quality evaluation—The case of Arabic, English, and Korean Wikipedias. *Library & information science research* 31, 4 (2009), 232–239.

[17] Besiki Stvilia, Michael B Twidale, Les Gasser, and Linda C Smith. 2005. Information quality discussions in Wikipedia. In *Proceedings of the 2005 International Conference on Knowledge Management.* 101–113.

[18] Besiki Stvilia, Michael B. Twidale, Linda C. Smith, and Les Gasser. 2005. Assessing Information Quality of a Community-based Encyclopedia. In *Proceedings of the International Conference on Information Quality.* 442–454. DOI: http://dx.doi.org/10.1.1.78.6243

[19] Besiki Stvilia, Michael B. Twidale, Linda C. Smith, and Les Gasser. 2008. Information quality work organization in wikipedia. *Journal of the American Society for Information Science and Technology* 59, 6 (2008), 983–1001. DOI: http://dx.doi.org/10.

1002/asi.20813 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.20813

[20] Morten Warncke-Wang. 2015. English Wikipedia Quality Asssessment Dataset. (4 2015). DOI: http://dx.doi.org/10.6084/m9.figshare.1375406.v1

[21] Morten Warncke-Wang, Vladislav R Ayukaev, Brent Hecht, and Loren G Terveen. 2015. The success and failure of quality improvement projects in peer production communities. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 743–756.

[22] Morten Warncke-Wang, Dan Cosley, and John Riedl. 2013. Tell Me More: An Actionable Quality Model for Wikipedia. In *Proceedings of the 9th International Symposium on Open Collaboration (WikiSym '13)*. ACM, New York, NY, USA, Article 8, 10 pages. DOI: http://dx.doi.org/10.1145/2491055.2491063

[23] Dennis M Wilkinson and Bernardo A Huberman. 2007. Cooperation and quality in wikipedia. *Proceedings of the 2007 International Symposium on Wikis* (2007), 157–164. DOI: http://dx.doi.org/10.1145/1296951.1296968

[24] Kewen Wu, Qinghua Zhu, Yuxiang Zhao, and Hua Zheng. 2010. Mining the Factors Affecting the Quality of Wikipedia Articles. In *2010 International Conference of Information Science and Management Engineering*, Vol. 1. 343–346. DOI: http://dx.doi.org/10.1109/ISME.2010.114