


# Quality of Wikipedia Articles: Analyzing Features and Building a Ground Truth for Supervised Classification

Elias Bassani<sup>1,2</sup> and Marco Viviani<sup>1</sup> 

<sup>1</sup>University of Milano-Bicocca, Department of Informatics, Systems, and Communication,  
Edificio U14 - Viale Sarca, 336, 20126 Milan, Italy

<sup>2</sup>Consorzio per il Trasferimento Tecnologico (C2T), Milan, Italy

**Keywords:** Data Quality, Wikipedia, Supervised Classification, Feature Analysis, Ground Truth Building.

**Abstract:** Wikipedia is nowadays one of the biggest online resources on which users rely as a source of information. The amount of collaboratively generated content that is sent to the online encyclopedia every day can let to the possible creation of low-quality articles (and, consequently, misinformation) if not properly monitored and revised. For this reason, in this paper, the problem of automatically assessing the quality of Wikipedia articles is considered. In particular, the focus is (i) on the analysis of groups of hand-crafted features that can be employed by supervised machine learning techniques to classify Wikipedia articles on qualitative bases, and (ii) on the analysis of some issues behind the construction of a suitable ground truth. Evaluations are performed, on the analyzed features and on a specifically built labeled dataset, by implementing different supervised classifiers based on distinct machine learning algorithms, which produced promising results.


## 1 INTRODUCTION

Web 2.0 technologies have given everyone the chance to generate and spread content online, in most cases without the intermediation of any traditional authoritative entity in charge of content control (Eysenbach, 2008; Ferrari and Viviani, 2013; Viviani and Pasi, 2017b). This augments the probability for people to incur into *misinformation* (Viviani and Pasi, 2017a), or *low-quality information* (Batini and Scannapieco, 2016). Online, traditional methods to estimate information quality – such as the scrupulous analysis of contents by experts – have become impractical, due to the huge amount of new content that is generated and shared every day on the Web. Therefore, it is necessary to design scalable and inexpensive systems to automatically estimate the quality of the information diffused, based on ‘objective’ evidence.

One of the main sources of knowledge freely accessible and editable by users, today, is Wikipedia.<sup>1</sup> The peculiarity of the platform, i.e., the fact that it allows anyone to create and modify articles, constitutes both a strength and a weakness: on the one hand, this encourages the collaborative construction of knowl-

edge, but, on the other hand, this can lead to the possible generation of low-quality or biased articles. To overcome this problem, groups of volunteers periodically monitor the content of Wikipedia articles, but their limited number confronted with the articles growth rate do not allow an overall and constant control. Furthermore, the subjectivity connected to human assessors results in a different quality evaluation for different articles belonging to distinct topic areas.

In this context, the work described in this paper aims at automating the classification of Wikipedia articles on qualitative bases, by employing supervised learning. The article focuses, on the one hand, on the analysis of *groups of hand-crafted features* that can be employed by well-known machine learning techniques (some of which previously applied in the literature) to classify Wikipedia articles over quality classes (an in-depth analysis has been performed both on the syntax, the style and the editorial history of Wikipedia articles). On the other hand, it focuses on the study of the Wikipedia classification process, highlighting very relevant aspects not previously treated in the literature in the construction of a suitable *ground truth*. For evaluation purposes, a specifically built labeled dataset has been generated, which is made publicly available. The results obtained by considering the analyzed features and ground truth

<sup>a</sup>  <https://orcid.org/0000-0002-2274-9050>

<sup>1</sup><https://www.wikipedia.org/>

confirm the effectiveness and the utility of the study.

## 2 BACKGROUND: WIKIPEDIA

Wikipedia is a collaborative encyclopedia where users can directly create new articles or modify them online. Due to this collaborative nature, the administrators of the platform (around 1,200 for the English version of Wikipedia) are required to constantly monitor the quality of the contents generated. Face to the huge flow of new information that everyday characterizes the publishing activity on Wikipedia, a manual monitoring is practically impossible. This leads to the introduction into the platform of a huge number of just sketched or low-quality articles. To cope with this problem and to indicate the *qualitative status* of an article, the Editorial Team of Wikipedia has defined some characteristics that an article should have in order to be considered of good quality, and distinct *quality classes* in which each article can be categorized based on its characteristics. Groups of contributors, called *WikiProjects*, are focused on improving the articles belonging to particular topic areas (e.g., *Mathematics*, *History*, etc.). Within each WikiProject, a so-called *assessment team* deals with the evaluation of the quality of the articles, which relies on the WikiProject article quality grading scheme.<sup>2</sup> It divides the articles into seven distinct categories: (i) *Featured Articles*, denoted as FA-Class (FA) articles; (ii) *A-Class (A)* articles; (iii) *Good Articles*, denoted as GA-Class (GA) articles; (iv) *B-Class (B)* articles; (v) *C-Class (C)* articles; (vi) *Start-Class (Start)* articles; (ii) *Stub-Class (Stub)* articles. The FA-Class includes the best articles on the platform, i.e., those considered complete and exhaustive from every point of view. In contrast, the Stub-Class includes all those articles that have a very basic description of the topic they deal with, or which are of very low quality. Intermediate classes are quality decreasing compared to the order in which they were previously listed.

## 3 RELATED WORK

Over the years, several works tackling the problem of *automatically* classifying Wikipedia articles with respect to the above-mentioned quality classes have been proposed. First approaches employing machine learning algorithms to perform classification, inferred evidence of the quality of articles only by considering *text features*, i.e., features connected to the length

<sup>2</sup>[https://en.wikipedia.org/wiki/Template:Grading\\_scheme](https://en.wikipedia.org/wiki/Template:Grading_scheme)

of the text (Blumenstock, 2008), the language usage (Lipka and Stein, 2010), or some lexical aspects (Xu and Luo, 2011). Other works have proposed *graph-based models* to estimate a quantitative value representing the quality of an article (Hu et al., 2007; Korfiatis et al., 2006; Li et al., 2015). These models consider and combine different metrics related to both: (i) the graph representing the editorial process of the articles, highlighting the relationships (edges) between articles and editors (nodes); (ii) the graph representing links (edges) among articles (nodes), i.e., the Wikipedia articles graph. In general, the models proposed within this group evaluate both authors authority and articles quality. Another category of approaches employs (supervised) machine learning techniques acting on *multiple kinds of features*, encompassing text and other features related to the writing style, the readability level, the analysis of the article structure, and other network-related metrics, to perform classification (Dalip et al., 2009; Dalip et al., 2014; Rassbach et al., 2007; Stvilia et al., 2005).

Recently, a few approaches based on the use of Deep Learning have been proposed (Dang and Ignat, 2016; Dang and Ignat, 2017). These approaches have proven to be effective in classifying Wikipedia articles over quality classes. Despite this, they do not involve *hand-crafted* feature analysis, while one of the aims of this paper is to study and investigate the impact that specific *groups of features* have in assessing information quality on Wikipedia.

## 4 FEATURE ANALYSIS

The choice of the features used to represent the elements on which to perform an automatic classification, is a fundamental operation in the majority of data-driven approaches (Fontanarava et al., 2017). In the literature, the number of features employed to automatically assessing the quality of Wikipedia articles was quite limited: the most complete study is (Dalip et al., 2009), in which 69 features were considered. The present work extends the number of features to 264. This consistent number derives, in large part, from an in-depth analysis of the *use of language* and the way in which sentences are constructed, and, therefore, the stylistic characteristics of the text. Furthermore, also the *editorial history* has been analyzed in depth, highlighting various aspects not yet considered in the literature, such as the contributions deriving from the changes made to the articles by *occasional users*. In the following, the considered features are detailed by regrouping them into *Text Features*, *Review Features*, and *Network Features*. Be-

cause of their high number, a synthetic description will be provided only when not self-evident. With respect to state-of-the-art features, new features introduced in this paper are indicated by an asterisk.

#### 4.1 Text Features

*Text Features* are extracted directly from the text of the articles. They allow to highlight, for example, the writing style, the structure, and the employed lexicon. For this reason, they can be further divided into four sub-categories: (i) **Length Features**, connected to some length aspects of the articles; (ii) **Structure Features**, capturing the way in which articles are structured (e.g., paragraphs); (iii) **Style Features**, highlighting the writing style and, therefore, the choices concerning the structuring of sentences and the use of the lexicon in drafting the articles; (iv) **Readability Features**, indicating the *degree of readability* of the articles, i.e., the minimum scholastic level that is necessary to understand their contents. In this paper, the study of new features focused on this area in particular, as the articles are mainly written texts and the textual characteristics turn out to be those that require less time and computational resources to be extracted. Moreover, these characteristics are, apart from rare cases, applicable in any context where there is a need to classify written texts on the basis of their quality, i.e., they are *platform-independent*.

**Length Features.** The length of an article can be an indicator of its quality. In fact, a good-quality text in a mature stage is reasonably neither too short (incomplete topic coverage), nor excessively long (verbose content). Further, in Wikipedia, *Stub* articles (draft quality) are short in the majority of the cases, reinforcing the correlation between length and quality (Dalip et al., 2009).

In this work, the following features have been considered: (1), (2), (3) **Character** (Stvilia et al., 2005) / **Word** (Rassbach et al., 2007) / **Sentence** (Dalip et al., 2009) **count**: the number of characters (including spaces) / words / sentences in the text; in addition, the new feature (4) **Syllable count**\* has been introduced, counting the number of syllables in the text.

**Structure Features.** This group of features focuses on the way an article is (well/badly) organized. According to the Wikipedia quality standards,<sup>3</sup> a good article must be reasonably clear, organized adequately, visually adequate, and point to appropriate references and/or external links.

<sup>3</sup>[https://en.wikipedia.org/wiki/Wikipedia:Version\\_1.0\\_Editorial\\_Team/Release\\_Version\\_Criteria](https://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Release_Version_Criteria)

In the following, the majority of state-of-the-art features come from (Dalip et al., 2009), unless otherwise indicated: (1), (2) **Section / Subsection count**: the number of sections / subsections in the article. The intuition behind these features is that a good article is organized in sections (e.g., *Introduction*, *Summary*, *List of references*, and *External links*) and subsections; (3) **Paragraph count**\*: the number of paragraphs constituting the article. The intuition behind this feature is that, in a high quality article, the text of sections and subsections should be further subdivided to facilitate the operation of reading and understanding the topics covered; (4), (5) **Mean section / paragraph size**; (6), (7) **Size of the longest / shortest section**, expressed in characters; (8) **Longest-Shortest section ratio**\*. This feature is useful to detect unusual section organization of articles with empty or very small sections, which could indicate incomplete content and drafts; (9) **Standard deviation of the section size**; (10) **Mean of subsections per section**; (11) **Abstract size**, expressed in characters. Mature articles are expected to have an introductory section summarizing its content; (12) **Abstract size-Article Length ratio**\*: an article presenting an abstract whose length is very similar to its total length is probably incomplete. Features (4) – (12) focuses on the correct balancing of an article.

Other structure features are: (13) – (15) **Citation count / count per section / count per text length**: the number of citations in the article/in sections/with respect to the total length of the article. A good-written article provides a sufficient and balanced number of citations; (16) – (18) **External link count** (Stvilia et al., 2005) / **links per section / links per text length**: the same rationale behind features (13) – (15); (19), (20) **Image count** (Stvilia et al., 2005) / **Images per section**: the number of images in the text and the ratio between the number of images and sections. Pictures contribute to make content clearer and visually pleasant; (21) **Images per text length**\*: the ratio between the number of images and the length of the article, expressed in number of sentences.

**Style Features.** Aim of these features is to capture the writing style of contributors, i.e., “some distinguishable characteristics related to the word usage, such as short sentences” (Dalip et al., 2009). Many of the style features reported here below have been employed in (Dalip et al., 2009). When possible, for each feature the first work having proposed it will be also indicated.

The considered style features are: (1), (2), (3) **Mean** (Xu and Luo, 2011) / **Largest** (Rassbach et al., 2007) / **Shortest**\* **sentence size**: the average num-

ber of words per sentence / the number of words of the longest / shortest sentence; (4), (5) **Large / Short sentence rate** (Rassbach et al., 2007): the percentage of sentences whose length is ten words greater / five words lesser than the article average sentence length; (6), (7) **Question count** (Rassbach et al., 2007) / **ratio\***: the number of questions in the article / the ratio between question count and the total number of sentences in the article; (8), (9) **Exclamation count\* / ratio\***; (10) – (17) **Number of sentences that start with a pronoun / an article / a coordinating conjunction / subordinating preposition or conjunction** (Rassbach et al., 2007) / **a determiner\* / an adjective\* / a noun\* / an adverb\***; (18) – (25) **Number of sentences that start with a pronoun- / an article- / a coordinating conjunction- / a subordinating preposition or conjunction- / a determiner- / an adjective- / a noun- / an adverb-** **Sentence count ratio\***: these features are built by considering the ratio between the value of features (10) – (17) and the total number of sentences that make up the article; (26) – (44) **Number of modal auxiliary verbs\* / passive voices** (Rassbach et al., 2007) / **‘to be’ verbs\* / different words\* / nouns\* / different nouns\* / verbs\* / different verbs\* / pronouns** (Dalip et al., 2009) / **different pronouns\* / adjectives\* / different adjectives\* / adverbs\* / different adverbs\* / coordinating conjunctions\* / different coordinating conjunctions\* / subordinating prepositions and conjunctions\* / different subordinating prepositions and conjunctions\*** in the whole article.

Features (45) – (62), which is a whole new group of features with respect to the literature, are the same as the group (26) – (44) but computed per each sentence constituting the article.

(63) – (80) **Ratio between the number of modal auxiliary verbs\* / passive voices\* / ‘to be’ verbs** (Rassbach et al., 2007) / **different words** (Xu and Luo, 2011) / **nouns** (Xu and Luo, 2011) / **different nouns\* / verbs** (Xu and Luo, 2011) / **different verbs\* / pronouns\* / different pronouns\* / adjectives\* / different adjectives\* / adverbs\* / different adverbs\* / coordinating conjunctions\* / different coordinating conjunctions\* / subordinating prepositions and conjunctions\* / different subordinating prepositions and conjunctions\*** and the total number of words in the article; (80) – (82) **Ratio between the number of modal auxiliary verbs\* / passive voice count\* / ‘to be’ verb\* and the total number of verbs in the article**; (83) – (89) **Ratio between the number of different nouns** (Xu and Luo, 2011) / **different verbs** (Xu and Luo, 2011) / **different pronouns\* / different adjectives\* / different adverbs\* / different coordinating conjunctions\* /**

**different subordinating prepositions and conjunctions\* and the total number of different words in the article**; (90) – (91) **Average number of syllables / characters per words**; (92) **Top- $m$  most discriminant character trigrams** (Lipka and Stein, 2010): they unveil the preferences of the authors for sentence transitions, as well as the utilization of stop-words, adverbs, and punctuation; (93) **Top- $n$  most discriminant POS trigrams**: they unveil the preferences of authors in constructing sentences (Lipka and Stein, 2010). To compute  $m$  and  $n$  for features (92) and (93), the  $\chi^2$  statistical method provided by the Python library `scikit-learn` has been employed.<sup>4</sup>

**Readability Features.** They are numerical indicators of the *US grade level*, i.e., the comprehension level that a reader must possess to understand what is debated in a text. They were first used, to tackle the considered problem, in (Rassbach et al., 2007).

The set of considered features includes: (1) **Automated Readability Index** (Smith and Senter, 1967); (2) **Coleman-Liau Index** (Coleman and Liau, 1975); (3) **Flesch Reading Ease** (Flesch, 1948); (4) **Flesch-Kincaid Grade Level** (Ressler, 1993); (5) **Gunning Fog Index** (Gunning, 1952); (6) **Läsbarhets Index** (Björnsson, 1968); (7) **SMOG-Grade** (Mc Laughlin, 1969). In this work, also (8) **Dale-Chall Readability Formula\*** (Chall and Dale, 1995) has been investigated. This latter metric, not previously employed for the quality assessment of Wikipedia articles, has been designed to numerically evaluate the difficulty of understanding that a reader encounters when s/he reads a text in English.

## 4.2 Review Features

*Review features* are extracted from the review history of each article, i.e., how many times and in which way the article has been modified. They can measure the degree of maturity and stability of an article, since no extensive corrections could indicate good-quality articles having reached a maturity level, while a lack of stability could indicate different kinds of controversies (e.g., with respect to neutrality, correctness, etc.). In the following list of review features, *registered users* are those having an explicit user profile and a username, *anonymous users* (Damiani and Viviani, 2009) are those identified only by their IP address, and *occasional users* are those who edited the article less than four times (they may belong to one of the two categories mentioned above). The considered features are: (1) **Age** (Rassbach et al., 2007):

<sup>4</sup>[http://scikit-learn.org/stable/modules/generated/sklearn.feature\\_selection.chi2.html](http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html)



the age (in days) of the article. Very recent articles are not normally considered of very high quality since they usually go through a refinement process; (2) **Age per review** (Dalip et al., 2009): the ratio between age and number of reviews. It is used to verify the average period of time an article remains without revision; (3) **Review per day** (Dalip et al., 2009): the percentage of reviews per day, to verify how frequently the article has been reviewed; (4) **Reviews per user** (Dondio et al., 2006): the ratio between the number of reviews and the number of users. This feature is useful to infer how much reviewed is an article when contrasted against the number of reviewers; (5) **Reviews per user standard deviation** (Dondio et al., 2006): this feature is useful to infer how balanced is the reviewing process among the reviewers; (6) **Discussion count** (Dondio et al., 2006): the number of discussions posted by the users about the article. This is useful to infer conflict resolution and teamwork dynamics; (7) **Review count** (Lih, 2004): the total number of reviews. (8) **User count** (Lih, 2004): the total number of unique users that have contributed to the article. More contributors an article has, more objective its content is supposed to be; (9) – (11) **Registered\* / anonymous\* / occasional\* user count**; (12) – (14) **Registered / anonymous / occasional user rate\***: Percentage of registered / anonymous / occasional contributors; (15) **Registered/Anonymous user ratio\***: the ratio between registered and anonymous contributors; (16) – (18) **Registered (Stvilia et al., 2005) / anonymous (Stvilia et al., 2005) / occasional\* review count** (Stvilia et al., 2005): the number of reviews made by registered / anonymous / occasional users; (19 – 21) **Registered\* / anonymous\* / occasional (Dondio et al., 2006) review rate\***: the percentage of reviews made by registered / anonymous / occasional users; (22) **Registered-Anonymous review ratio\***: the ratio between reviews made by registered users and anonymous users.

Features (9) – (22), previously unconsidered in the literature, aim at highlighting the possible qualitative difference between articles based on the role that in the Wikipedia platform have registered, anonymous and occasional users.

(23) **Revert count** (Stvilia et al., 2005): the number of times an article has been taken to a previous state (review annulment); (24) **Reverts count-Review count ratio\***: the ratio between reverts count and review count; (25) **Diversity** (Stvilia et al., 2005): the ratio between the total number of contributors and the number of reviews; (26) **Modified lines rate** (Dalip et al., 2009): the number of lines modified when comparing the current version of an article with

three-months older version. This is a good indicator of how stable an article is; (27) **Last three-months review count\***: the number of reviews made in the last three months. This feature could indicate that the content of an article is controversial, the article is about evolving events or it is in the beginning of its editorial process; (28) **Last three-months review rate** (Dondio et al., 2006): the percentage of reviews made in the last three months; (29) **Most active users review count\***: the number of reviews made by the most active 5% of users; (30) **Most active users review rate** (Dondio et al., 2006): the percentage of reviews made by the most active 5% of users; (31) **ProbReview** (Hu et al., 2007): this measure tries to assess the quality of a Wikipedia article based on the quality of its reviewers. Recursively, the quality of the reviewers is based on the quality of the articles they reviewed.

### 4.3 Network Features

*Network features* are extracted from the articles graph, which is built by considering citations among articles. These citations can provide evidences of the popularity of the articles. A high-quality article is expected to be used as a reference point for articles dealing with interconnected topics. Extracting this kind of feature is particularly onerous, due the magnitude of the graph.

For this reason, state-of-the-art features have been considered: (1) **PageRank** (Brin and Page, 2012): the *PageRank* of an article, previously employed in (Rassbach et al., 2007); (2) **In-degree** (Dalip et al., 2009): the number of times an article is cited by other articles; (3) **Out-degree** (Dondio et al., 2006): the number of citations of other articles; (4) – (7) **Assortativity in-in / in-out / out-in / out-out**: “the ratio between the degree of the node and the average degree of its neighbors. The degree of a node is defined as the number of edges that point to it (in-degree) or that are pointed by it (out-degree)” (Dalip et al., 2009); (8) **Local clustering coefficient** (Watts and Strogatz, 1998): it aims at evaluating if an article belongs to a group of correlated articles.

Features (4) – (8), related to assortativity and clustering coefficient were proposed in (Benevenuto et al., 2008; Castillo et al., 2007; Dorogovtsev and Mendes, 2013) for *spam detection* in Web pages. They were previously used in (Dalip et al., 2009).

(9) **Reciprocity**: the ratio between the number of articles that cite a specific article and the number of articles cited by that article; (10) **Link count** (Dondio et al., 2006): the number of links to other articles. It differs from *out-degree* since it counts also links to ar-

ticles that have not been written yet (*red links*);<sup>5</sup> (11) **Translation count** (Dondio et al., 2006): the number of versions of an article in other languages.

## 5 BUILDING A GROUND TRUTH

In supervised approaches for classification, the availability of a labeled dataset is a mandatory condition to train the proposed classifiers. In the literature, prior supervised approaches in the context of Wikipedia article quality assessment have employed datasets that were labeled in different ways. In fact, depending on their aims, previous works classified the articles of Wikipedia with respect to a *subset* of the seven quality classes considered in this article (illustrated in Section 2). Specifically, none of the previous studies have referred to the current grading scale, for different reasons: (i) when the study was made the proposed quality scale was different (Dalip et al., 2009); (ii) the authors decided to simplify the classification task on purpose. In (Rassbach et al., 2007) the Stub-Class (drafts) has not been considered because it was believed to be too trivial to discern articles belonging to that class. In (Xu and Luo, 2011) the authors consider only *Featured Articles* and Start-Class articles. In (Blumenstock, 2008; Lipka and Stein, 2010; Xu and Luo, 2011) the articles were classified only as *Featured Articles* and *Random Articles* (or *non-Featured Articles*); (iii) the approaches did not perform a classification into quality classes, but, rather, a ranking of articles with respect to their quality. The ranking produced by (Hu et al., 2007) is supposed to reflect the hierarchy of the quality classes: “the perfect ranking should place all FA-Class articles before all A-Class articles, followed by all GA-Class articles and so on”, while in (Li et al., 2015) the ranking is intended to identify *relevant VS non-relevant* articles, i.e., *Featured VS non-Featured Articles*.

### 5.1 Multi-class Classification

Aim of this study, is to allow to automatically perform a so-called (single-label) *multi-class* classification, where each article is assigned exactly to one of the seven quality classes that Wikipedia employs nowadays. Therefore, with respect to the approaches presented in the literature, it has been necessary to proceed with the construction of a new dataset, by selecting labeled articles (from the seven quality classes detailed in Section 2) directly from Wikipedia. In doing so, some aspects not discussed in detail in previous works have been considered.

<sup>5</sup>[https://en.wikipedia.org/wiki/Wikipedia:Red\\_link](https://en.wikipedia.org/wiki/Wikipedia:Red_link)

First of all, the guidelines for the classification provided by the Wikipedia Editorial Team are generic, and need to be specialized/refined according to the topic area considered. For example, a high-quality article discussing *Photography* is expected to have more images than one of similar quality dealing with *Computer Science*, while a *History* article is supposed to contain more dates with respect to *Technology* ones, which probably will contain more technical details. The linguistic register is another aspect that it is influenced by thematic areas: a high-quality *Economy* article will be characterized by a more complex lexicon compared to the one employed in a *Literature for children* article. Moreover, an article can be provided with multiple quality labels by distinct WikiProjects if its content touches different topic areas.

In most of the works proposed in the literature, labeled datasets containing articles belonging to *multiple topic areas* have been employed together, thus not dealing with the above-described issues (Blumenstock, 2008; Xu and Luo, 2011; Stvilia et al., 2005; Rassbach et al., 2007; Dalip et al., 2009). These approaches focus on the development of supervised techniques able to classify each Wikipedia article with respect to quality classes disregarding its topic area (and, therefore, the WikiProject it belongs to). Other approaches have considered *specific topic areas* to build suitable dataset to train their models (one labeled dataset per area), in order to provide different classifications per each WikiProject (Hu et al., 2007; Lipka and Stein, 2010; Li et al., 2015). In these works, the choices undertaken in the selection of specific topic areas were not detailed.

In the same spirit of considering distinct areas, in this paper the idea is to provide WikiProject teams with an effective instrument able to train the quality classifier with respect to the area of interest. In this way, the features remain those defined in Sections 4.1, 4.2, and 4.3; what changes are the articles that make up each labeled dataset to be constructed.

### 5.2 A Specific Dataset

To experimentally evaluate the effectiveness of the analyzed features (Section 4) in the context of (single-label) multi-class classification, a first labeled dataset, namely Dataset (a), has been constructed for the topic area *Military History*,<sup>6</sup> which was, at the time of writing, the one with the highest number of articles per quality class, i.e., 1,090 FA-Class, 564 A-Class, 4,049 GA-Class, 13,675 B-Class, 22,604 C-Class, 77,908 Start-Class, and 50,281 Stub-Class articles. The obtained dataset consists of 400 articles randomly se-

<sup>6</sup>[https://en.wikipedia.org/wiki/Category:Military\\_history](https://en.wikipedia.org/wiki/Category:Military_history)

lected for each quality class, for a total of 2,800 articles. The choice of considering this number of articles derives by the small size of the minority class (i.e., the A-Class); by randomly selecting for each class the same number of articles, we are able to act on a *balanced* dataset.

During the construction of the dataset, an aspect worthy of consideration has emerged, which apparently had not been previously taken into consideration in the literature. As said before, Wikipedia is a highly dynamic platform whose contents are constantly modified. Therefore, the possibility that the version of the articles constituting Dataset (a) - recently gathered from Wikipedia - is not the same version on which the original classification made by the Military History WikiProject was performed, is far from remote. In fact, after an in-depth analysis, it has emerged that the articles present in Dataset (a) were classified a significant amount of time before (and many versions before) the recently gathered version. For this reason, a second dataset, namely Dataset (b), has been built, composed of the same articles as Dataset (a), but containing the original classified versions of the articles, i.e., the texts based on which the classification was performed. Both datasets are made publicly available, together with instructions on their usage.<sup>7</sup>

## 6 EXPERIMENTAL EVALUATION

In this section, the effectiveness of the considered features is evaluated by testing different supervised machine learning classifiers, by employing features detailed in Section 4, and the labeled datasets described in Section 5.2. Specifically, *four experiments* are illustrated, aiming at evaluating different aspects, as detailed in the dedicated sections. Each experiment tests eight different classifiers based on distinct supervised machine learning techniques: *Decision Tree* (DT), *K-Nearest Neighbors* (KNN), *Logistic Regression* (LR), *Naive Bayes* (NB), *Random Forest* (RF), *Support Vector Classifier* (SVC), *Neural Networks* (NN) and *Gradient Boosting* (GB). The classifiers have been implemented by using the Python *Scikit-learn* library,<sup>8</sup> in particular for the first seven classifiers, and the *XGBoost* library<sup>9</sup> for the last classifier. In each experiment, a *k*-fold cross-validation has been performed, where *k* = 20; the classification performance has been evaluated in terms of *Accuracy* and *Mean-Squared Error* (MSE) (Kubat, 2015).

<sup>7</sup><https://github.com/ir-laboratory/wikipediadataset>

<sup>8</sup><https://scikit-learn.org/>

<sup>9</sup><https://xgboost.readthedocs.io>

**Experiment 1.** In the first experiment, a comparative evaluation has been performed between the proposed approach and the state-of-the-art baseline [A] described in (Dalip et al., 2009), which employed supervised classifiers acting on the higher number of hand-crafted features among prior works in the literature. This experiment allows to evaluate the effectiveness of the features analyzed in this paper and those proposed by the baseline in classifying Wikipedia articles with respect to the seven quality classes (i.e., FA-Class, A-Class, GA-Class, B-Class, C-Class, Start-Class, and Stub-Class).

Table 1: Experiment 1 - Accuracy (higher is better) and MSE (lower is better).

Classifier	Proposed Approach		Baseline [A]	
	Acc.	MSE	Acc.	MSE
<b>DT</b>	0.47	1.77	0.48	1.77
<b>KNN</b>	0.42	2.12	0.42	2.06
<b>LR</b>	0.50	1.36	0.50	1.41
<b>NB</b>	0.30	3.57	0.31	3.44
<b>RF</b>	0.59	1.17	0.60	1.07
<b>SVC</b>	0.51	1.36	0.54	1.43
<b>NN</b>	0.50	1.20	0.49	1.35
<b>GB</b>	<b>0.62</b>	<b>0.92</b>	<b>0.60</b>	<b>1.03</b>

As reported in Table 1, the set of features analyzed in this paper in conjunction with Gradient Boosting allow to obtain the best results in terms of both Accuracy (62%) and MSE (0.92), with an improvement in terms of both measures with respect to [A].

**Experiment 2.** The second experiment consists in the classification of Wikipedia articles w.r.t. the seven quality classes by considering, in turn, only the features belonging to each of the three groups in which they can be categorized, i.e., *Text Features* (TF), *Review Features* (RF) and *Network Features* (NF). This experiment aims to identify which group of features is the most discriminating one in terms of article quality. In Table 2, the comparison between the accuracy and MSE values obtained with respect to each group of features are reported respectively.

Table 2: Experiment 2 - Accuracy and MSE.

Classifier	Accuracy			MSE		
	TF	RF	NF	TF	RF	NF
<b>DD</b>	0.38	0.32	0.30	2.08	3.01	3.61
<b>KNN</b>	0.42	0.29	0.30	2.10	4.40	3.96
<b>LR</b>	0.47	0.39	0.33	1.48	2.80	3.58
<b>NB</b>	0.30	0.25	0.20	3.47	6.41	9.05
<b>RF</b>	0.50	0.39	0.37	1.35	2.20	2.79
<b>SVC</b>	0.46	0.38	0.33	1.52	3.16	3.85
<b>NN</b>	0.48	0.39	0.37	1.28	2.66	2.85
<b>GB</b>	<b>0.51</b>	0.39	0.35	<b>1.17</b>	2.18	2.97

As it emerges from the table, *Network Features* appear to be the less effective, while *Text Features* provide the best level of Accuracy and MSE by employing Gradient Boosting. In particular, TF+GP provides an Accuracy value of 51%, and an MSE value of 1.17.

**Experiment 3.** As illustrated in Section 3, some prior works took in consideration the classification of Wikipedia articles only with respect to the two classes *Featured Articles* and *non-Featured Articles* (Blumenstock, 2008; Lipka and Stein, 2010). In this experiment, the proposed approach is therefore comparatively evaluated with respect to these two baselines, denoted as [B] and [C], in performing this binary classification. To do this, the dataset has been reduced to 800 articles, 400 *Featured Articles* and 400 *non-Featured Articles*, uniformly distributed to deal with a balanced set. As reported in Table 3, the considered features in conjunction with Gradient Boosting obtained the best results in terms of both Accuracy (90%) and MSE (0.09) also with respect to binary classification.

Table 3: Experiment 3 - Accuracy and MSE.

Cl.	P. Approach		[B]		[C]	
	Acc.	MSE	Acc.	MSE	Acc.	MSE
DD	0.84	0.16	0.72	0.28	0.78	0.22
KNN	0.79	0.21	0.79	0.20	0.79	0.21
LR	0.83	0.17	0.74	0.26	0.80	0.20
NB	0.71	0.29	0.68	0.32	0.68	0.32
RF	0.87	0.12	0.72	0.28	0.83	0.17
SVC	0.84	0.16	0.72	0.28	0.79	0.21
NN	0.85	0.15	0.50	0.50	0.78	0.22
GB	<b>0.90</b>	<b>0.09</b>	0.76	0.24	0.83	0.17

**Experiment 4.** In the last experiment, the impact of the noise introduced by the changes occurred in the articles after their classification by WikiProjects in the process of automatically detecting their quality class is evaluated. Specifically, the eight considered classifiers are trained over Dataset (a) and Dataset (b) described in Section 5.2. In particular, in this experiment, only a subset of the features presented in Section 2 has been employed, i.e., not considering *Network Features*, which have demonstrated, as detailed in Section 6, to be the less effective. This decision has been also taken to overcome some relevant technical issues arising from the impossibility to build a unique graph representing the versions of the articles present in dataset (b) in the exact moment they were classified, because (i) they were not all classified at the same time by the *Military History* WikiProject, and (ii) old link information between articles are not available - their dumps are deleted by Wikipedia if they are older than 6 months. The *ProbReview* fea-

ture was also not taken into account for this particular experiment due to its high computational cost. However, since the purpose of this experiment is to provide some insights about the effect of the introduction of noise on the automation of the classification task, this appeared to be a reasonable compromise.

As reported in Table 4, the best results with all the machine-learning-based classifiers have been reached with Dataset (b) with a remarkable margin. This experiment clearly demonstrates that the noise present in the first dataset has a non-negligible impact on the automation of the classification process and therefore it should be kept in mind in future studies.

Table 4: Experiment 4 - Accuracy and MSE.

Classifier	Dataset (a)		Dataset (b)	
	Acc.	MSE	Acc.	MSE
DD	0.375	2.002	0.452	1.570
KNN	0.418	2.164	0.435	1.857
LR	0.487	1.359	0.528	1.215
NB	0.302	3.512	0.341	2.791
RF	0.525	1.234	0.562	1.059
SVC	0.487	1.452	0.523	1.274
NN	0.494	1.174	0.529	1.067
GB	0.530	1.048	<b>0.584</b>	<b>0.912</b>

## 7 CONCLUSIONS

In this paper, the problem of automatically assessing the quality of Wikipedia articles has been considered, to combat the proliferation of unverified and low-quality contents. In the last years, several approaches for the classification of Wikipedia contents with respect to given quality classes have been proposed. Many of these solutions are based on supervised learning techniques, employing multiple kinds of features connected to different aspects of the articles and their authors.

With respect to state-of-the-art approaches, in this paper, the analysis of a higher number of hand-crafted features has been proposed. The choice of the features is based on an in-depth analysis that encompass the syntax, the style and the editorial history of Wikipedia articles, as well as on a deep investigation of the way in which Wikipedia articles are labeled by WikiProject teams with respect to quality. Furthermore, some investigations on how to build a suitable ground truth for the considered issue has been presented. The promising results obtained confirm the effectiveness of the proposed feature analysis and the interest in continuing the study of the problem, investigating some aspects connected to statistical significance, not addressed in this paper, and the comparison with Deep Learning approaches.



## REFERENCES

- Batini, C. and Scannapieco, M. (2016). *Data and Information Quality - Dimensions, Principles and Techniques*. Data-Centric Systems and Applications. Springer.
- Benevenuto, F., Rodrigues, T., Almeida, V. A. F., Almeida, J. M., Zhang, C., and Ross, K. W. (2008). Identifying video spammers in online social networks. In *AIRWeb 2008, China, April 22, 2008, Proc.*, pages 45–52.
- Björnsson, C.-H. (1968). *Lesbarkeit durch Lix*. Pedagogiskt centrum, Stockholms skolförvaltn.
- Blumenstock, J. E. (2008). Size matters: word count as a measure of quality on wikipedia. In *WWW 2008, Proceedings, Beijing, China, April 21-25, 2008*, pages 1095–1096.
- Brin, S. and Page, L. (2012). Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 56(18):3825–3833.
- Castillo, C., Donato, D., Gionis, A., Murdock, V., and Silvestri, F. (2007). Know your neighbors: web spam detection using the web topology. In *SIGIR 2007, Proceedings, Amsterdam, The Netherlands, July 23-27, 2007*, pages 423–430.
- Chall, J. S. and Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula*. Brookline Books.
- Coleman, M. and Liao, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.
- Dalip, D. H., Gonçalves, M. A., Cristo, M., and Calado, P. (2009). Automatic quality assessment of content created collaboratively by web communities: a case study of wikipedia. In *JCDL 2009, Proceedings, Austin, TX, USA, June 15-19, 2009*, pages 295–304.
- Dalip, D. H., Lima, H., Gonçalves, M. A., Cristo, M., and Calado, P. (2014). Quality assessment of collaborative content with minimal information. In *IEEE/ACM JCDL 2014, London, United Kingdom, September 8-12, 2014*, pages 201–210.
- Damiani, E. and Viviani, M. (2009). Trading anonymity for influence in open communities voting schemata. In *2009 International Workshop on Social Informatics*, pages 63–67. IEEE.
- Dang, Q. V. and Ignat, C.-L. (2016). Quality assessment of wikipedia articles without feature engineering. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, pages 27–30. ACM.
- Dang, Q.-V. and Ignat, C.-L. (2017). An end-to-end learning solution for assessing the quality of wikipedia articles. In *Proceedings of the 13th International Symposium on Open Collaboration*, page 4. ACM.
- Dondio, P., Barrett, S., and Weber, S. (2006). Calculating the trustworthiness of a wikipedia article using dante methodology. In *IADIS eSociety conference, Dublin, Ireland*.
- Dorogovtsev, S. N. and Mendes, J. F. (2013). *Evolution of networks: From biological nets to the Internet and WWW*. OUP Oxford.
- Eysenbach, G. (2008). Credibility of health information and digital media: New perspectives and implications for youth. In *Digital Media, Youth, and Credibility*, pages 123–154. The MIT Press.
- Ferrari, E. and Viviani, M. (2013). Privacy in social collaboration. In Michelucci, P., editor, *Handbook of Human Computation*, pages 857–878. Springer.
- Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, 32(3):221.
- Fontanarava, J., Pasi, G., and Viviani, M. (2017). Feature analysis for fake review detection through supervised classification. In *Data Science and Advanced Analytics (DSAA), 2017*, pages 658–666. IEEE.
- Gunning, R. (1952). The technique of clear writing.
- Hu, M., Lim, E., Sun, A., Lauw, H. W., and Vuong, B. (2007). Measuring article quality in wikipedia: models and evaluation. In *Proc. of CIKM 2007, Lisbon, Portugal, November 6-10, 2007*, pages 243–252.
- Korfiatis, N., Poulos, M., and Bokos, G. (2006). Evaluating authoritative sources using social networks: an insight from wikipedia. *Online Information Review*, 30(3):252–262.
- Kubat, M. (2015). *An introduction to machine learning*, volume 681. Springer.
- Li, X., Tang, J., Wang, T., Luo, Z., and de Rijke, M. (2015). Automatically assessing wikipedia article quality by exploiting article-editor networks. In *ECIR 2015, Vienna, Austria, March 29 - April 2, 2015. Proceedings*, pages 574–580.
- Lih, A. (2004). Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource. *Nature*.
- Lipka, N. and Stein, B. (2010). Identifying featured articles in wikipedia: writing style matters. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, pages 1147–1148.
- McLaughlin, G. H. (1969). Smog grading—a new readability formula. *Journal of reading*, 12(8):639–646.
- Rassbach, L., Pincock, T., and Mingus, B. (2007). Exploring the feasibility of automatically rating online article quality. In *WikiMania 2007, Proceedings, Taipei, Taiwan*, page 66.
- Ressler, S. (1993). *Perspectives on electronic publishing - standards, solutions, and more*. Prentice Hall.
- Smith, E. A. and Senter, R. (1967). Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories (US)*, pages 1–14.
- Stvilia, B., Twidale, M. B., Smith, L. C., and Gasser, L. (2005). Assessing information quality of a community-based encyclopedia. In *ICIQ 2005, Cambridge, MA, USA, November 10-12, 2006*.
- Viviani, M. and Pasi, G. (2017a). Credibility in social media: opinions, news, and health information—a survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(5):e1209.
- Viviani, M. and Pasi, G. (2017b). Quantifier guided aggregation for the veracity assessment of online reviews. *International Journal of Intelligent Systems*, 32(5):481–501.
- Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of ‘small-world’ networks. *Nature*, 393(6684):440–442.
- Xu, Y. and Luo, T. (2011). Measuring article quality in wikipedia: Lexical clue model. In *Web Society (SWS), 2011*, pages 141–146. IEEE.