Bauhaus-Universität Weimar
Faculty of Media
Degree Programme Media Systems

Featured Article Identification in Wikipedia

Bachelor Thesis

Christian Fricke

1. Supervisor: Prof. Dr. Benno Stein
2. Supervisor: Prof. Dr. Charles Wüthrich
Advisor: Maik Anderka

Date of submission: September 28, 2012

Declaration of Authorship

I hereby ensure that the thesis at hand is entirely my own work, employing only the referenced media and sources.

.....

Weimar, September 28, 2012 Christian Fricke

Abstract

Over the last decade, a variety of quality indicators have been introduced in an effort to automatically assess the quality of content created by collaborative communities such as Wikipedia. The effectiveness of the indicators has been tested by means of classifying featured and non-featured articles. In this thesis, we provide a comprehensive summary of the article features found in the relevant literature. Furthermore, we analyze several of the presented classification algorithms on uniform datasets for a fair assessment of the respective performance. We compare our results to the ones presented in the various studies and show which prove to be the most effective.

Contents

1 Introduction **5**
1.1 Wikipedia Fundamentals 7
1.2 Automatic Quality Assessment 11

2 Retrieving Articles **14**
2.1 Data Organization in Wikipedia 14
2.2 Exporting Mechanisms in Wikipedia 17

3 Representing Articles **22**
3.1 Content Features 23
3.2 Structure Features 27
3.3 Network Features 29
3.4 History Features 31

4 Classifying Articles **34**
4.1 Classification Methodology 34
4.2 Experiment Conclusion 38

5 Conclusion **43**

A Appendix **44**
A.1 Complete Feature List 44
A.2 Reference Section Names 47

Bibliography **48**

1 Introduction

In January 2001, Wikipedia was launched as an experiment to boost the content production of the free-content and peer-reviewed encyclopedia, Nupedia, created by Jimmy Wales and Larry Sanger. The term, Wikipedia, was coined by the latter as a compound of *wiki* and encyclo*pedia*. A wiki is a type of collaborative website, which allows for easy content manipulation using a simplified markup language. The first wiki software was called WikiWikiWeb, originally invented by Ward Cunningham in 1995.

In contrast to conventional multi-authored encyclopedias, Wikipedia is freely available to anyone and collaboratively edited by volunteers. Almost every article can be altered by the reader as a registered or anonymous user. Furthermore, an article is not owned by its creator. There is no authority other than the community itself, whose editors are supposed to agree on the content and structure by consensus. After creating or editing an article, its revision is immediately accessible and thus may contain inaccuracies, ideological biases, or plain nonsense. Because of its openness, Wikipedia has been critiqued for the quality of writing and accuracy of information.

The dimensions used in assessing information quality are manifold. Depending on the context, the term information may be interpreted in various ways, similar to the perception of quality as subjective. Intrinsic dimensions (e.g., accuracy) are independent of the user's context, whereas contextual dimensions (e.g., relevancy) need to be assessed based on subjective preferences [1]. More generally, Juran [2] defines quality as user oriented "fitness for use", the main aspects of which are adding features that meet the user's needs and reducing defects. In regard to Wikipedia, this translates to policies and guidelines developed by the community to resolve conflicts and describe best practices.

The process of information quality assurance guarantees confidence that particular information meets some context specific quality requirement [3]. The analogous quality assessment of Wikipedia articles has been of interest to many researchers alike. Most studies are concerned with the classification of articles by

means of predefined quality measures, as provided by the encyclopedia's article quality grading scheme, which we will discuss in more detail in Chapter 3.

Featured pages and discussion pages are among the most notable introductions to accommodate the need for improving the quality of articles. Former are used as an indication of exceptional user-generated content (as defined by the user's needs), while latter serve as a platform for augmenting the quality of all content pages (reducing defects).

As of January 2012, the English Wikipedia comprises more than 3.8 million articles, of which approximately 0.1% carry the featured tag. In order to be considered as such, individuals or a group of contributors has to nominate an article as a candidate. A peer review process conducted by the editors decides on whether or not these meet the featured article criteria, which are continuously evolving. The many different contexts this form of information creation entails require constant negotiation among the editors and the community. Therefore, one could argue that the quality of information is indeed subjective and depends on the needs of the consumer. Automatically assessing said quality is no trivial task, and simple scalar measures are not useful in this context. The amount of sections or references, for instance, indicate no particular level of quality on their own. Even within the same quality class, the perceived quality varies from individual to individual.

How do featured articles in Wikipedia differ from the ones that are non-featured? We assume a significant gap in quality between these two, hence describe the problem as a classification task. An article is either featured, when it meets certain criteria and has been nominated as such, or non-featured. Many algorithms have been devised on that basis. However, each of them has been tested on separate subsets (or specific domains) of the whole Wikipedia corpus. No general conclusions can be drawn from the results of the experiments conducted in regard to the overall performance on randomly chosen articles.

Our contributions are twofold. We implement the most promising of the algorithms presented in the relevant literature and analyze their performance on uniform datasets based on up-to-date content from the English Wikipedia. Furthermore, we provide a comprehensive summary of the data organization in Wikipedia and a detailed description of the metrics used to assess the quality of its articles, as well as a framework to consistently evaluate a variety of different classification models. The outcome of the conducted experiments validates the findings of other studies in the field.

Table 1.1: Example of keywords and syntax used for headings, style attributes, and page links (left), as employed by the MediaWiki software to produce an HTML rendered layout (right).

```
== Information quality ==

'''Information quality''' (IQ) is a
term to describe the quality of the
content of [[information systems]].
It is often pragmatically defined
as: "The ''fitness for use'' of
the information provided."
```

Information quality

Information quality (IQ) is a term to describe the quality of the content of information systems. It is often pragmatically defined as: "The *fitness for use* of the information provided."

1.1 Wikipedia Fundamentals

Wikipedia, with its large user base and wide range of articles, often functions as a corpus for studying and developing models and algorithms that are concerned with the task of information quality assessment. It offers a variety of grading schemes, which are employed by groups of editors within collections of articles sharing a common topic, referred to as *WikiProjects*. This section contains a short introduction to the general structure of Wikipedia, in which we explore the fundamental components shared between wikis of that kind.

Powered by the MediaWiki software, Wikipedia provides means for editing its *pages*, which account for the majority of its content, in a fast and easy manner. It only requires a simple Web browser and no prior knowledge about the software (other than a few syntactic rules). Leuf and Cunningham [4] further elaborate on the essence of wikis, promoting meaningful topic associations between different pages by creating straightforward page links, and seeking to involve the visitor in an ongoing process of creation and collaboration. Table 1.1 illustrates the markup for headings, style attributes and page links in a nutshell. A detailed overview of all the keywords and syntactic rules can be found at the internal help website for wiki markup[1].

There are currently 22 different kinds of pages, each assigned a dedicated namespace[2]. The main namespace subsumes all pages with encyclopedic information, henceforth referred to as *articles*. These are commonly organized

[1] http://en.wikipedia.org/wiki/Help:Wiki_markup
[2] The concept of namespaces is further explored in Section 2.1.

in categories according to their subject matter and user-assessed quality status. The category system provides navigational links to all Wikipedia pages in form of a hierarchical taxonomy. Hence, topic related pages can be explored based on their defining characteristics. Lists and additional navigation boxes may also be used to connect relevant articles. Wikis achieve easy linking between pages within its domain by employing interwiki or (in the case of Wikipedia) interwikimedia links. Internal links share the same root URL, enabling users to avoid pasting entire URLs that would otherwise make fast editing rather problematic.

A user's contribution to an article is recorded as a revision, more commonly referred to as an *edit*. Preserved in a complementary history page, similar in function to a version control system, no edit is lost and can effortlessly be restored, for instance, in case of vandalism. Other means of editing include *templates* and the above-mentioned *discussion* pages, also known as talk pages. Each article is associated with a talk page, which editors may use to discuss changes, coordinate work or reach consensus on its quality. Templates, on the other hand, are pages to be included in other pages. These contain commonly used messages, warnings, lists or boxes. Most wikis facilitate a varying degree of search functionality. Wikipedia's built-in engine supports searching all its namespaces, defaulting to the mainspace.

In Wikipedia, a supplementary access layer prevents certain pages from being modified or damaged. The placing of so-called protections restricts editing permissions in varying degrees, ranging from full to function-specific preservation. These can only be applied and removed by Wikipedia's administrators, a user group that has been granted the technical ability to perform special actions on Wikipedia. Other groups include account creators, bots, file movers, and reviewers.

1.1.1 Article Grading Principles

The English Wikipedia employs a variety of policies and guidelines developed by the community that serve to document good practices. As stated in the "What Wikipedia is not" project page[3], Wikipedia is not a bureaucracy. The rules merely document already existing community consensus regarding the acceptance and rejection of content additions. The first in the list of policies, "Ignore all rules", reflects this view by stating only the following:

[3] http://en.wikipedia.org/wiki/Wikipedia:NOT

> If a rule prevents you from improving or maintaining Wikipedia, ignore it.

The most pertinent principles by which Wikipedia is governed are enumerated in the five pillars[4]. These are summarized as follows:

Encyclopedia Wikipedia is a project to create a freely available compendium which conveys information on human knowledge. It has been verbalized quite copiously what it is not, such as an anarchy, censored, for unverifiable material, limited by paper, a web directory or a newspaper, to name just a few examples.

Point of view The three core content policies are "Neutral point of view", "Verifiability", and "No original research". All articles are composed free from personal experience, interpretations or opinions. In the case of multiple points of view, editors present each one accurately and in context.

Free content Anyone can edit, use, modify, and distribute Wikipedia's content. There is no concept of authorship of an article; all contributions are freely licensed to the public.

Civility The *Wikiquette* describes behavioral guidelines for the many different editors, who should interact with each other in a respectful manner, avoid edit wars, and focus on the task of improving rather than disrupting Wikipedia.

No rules Wikipedia is not governed by statutes and has no firm rules. The literal wording of the policies and guidelines is less important than the general principles they try to convey. Rules may need to be broken in order to improve the quality of the content.

A multitude of studies are concerned with the implied open access to all of Wikipedia's articles [5, 6, 7, 8, 9, 10]. Such an editing philosophy can never ensure reliable and accurate content. The quality of traditional encyclopedias, written and published in a centralized manner, generally supersedes the one of those that are collaboratively edited, which lack further restrictions and quality control. Giles [11] presents the results from an expert-led investigation to compare the quality of randomly selected science articles from Wikipedia to those of the *Encyclopedia Britannica*. They identify an average number of four and three inaccuracies per reviewed article for each encyclopedia respectively,

[4] http://en.wikipedia.org/wiki/Wikipedia:Five_pillars

approaching Jimmy Wales' goal of "getting Britannica quality". In an appeal[5] in 2009, the Wikipedia co-founder communicates his vision of improving the quality through relentless contribution by the community:

> I believe that Wikipedia keeps getting better. That's the whole idea. One person writes something, somebody improves it a little, and it keeps getting better, over time. If you find it useful today, imagine how much we can achieve together in 5, 10, 20 years.

In an effort to create an offline release version of Wikipedia a group was formed in 2004 that would further identify and organize a core set of articles to be included in the final distribution. The *Version 1.0 Editorial Team* employ a bot-assisted selection process based on the manual quality assessment by WikiProjects. Currently 3 100 000 articles have been analyzed this way. The team maintains a grading system to evaluate an article's progression towards distribution quality. It serves as a guideline for members of WikiProjects who perform most of the classification. The *WP 1.0 Bot*[6] tracks talk page banners for individual projects and produces statistics by taking the highest quality and importance rating for each assessed article in the main namespace. Table 1.2 summarizes the grading scheme that is used to outline the various quality classes. A separate list of criteria for each class is used to manually assess an article's quality. Furthermore, independent panels manage the nominated candidates for categories GA and FA, the criteria for which can be found at the internal project websites for featured[7] and good[8], respectively. A *stub*, per definition, is an article too short to provide encyclopedic coverage of a subject and is generally marked by a stub template until it has been augmented with meaningful content. Stubs are typically excluded from the presented studies in the following section. The above system indicates the quality of an article as a value on a discrete scale. However, a continuous scale allows for a finer distinction between articles that share the same category, thus making suggestions for demotion and promotion possible [12].

[5] http://wikimediafoundation.org/wiki/Appeal/en
[6] A small application that automatically performs simple and structurally repetitive tasks.
[7] http://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria
[8] http://en.wikipedia.org/wiki/Wikipedia:Good_article_criteria

Table 1.2: Wikipedia's article quality grading scheme, containing associated color, class name and editing suggestions.

Class	Editing Suggestions
FA	no further additions necessary unless new information becomes available
A	expert knowledge may be required to further improve the article
GA	revision by subject and style experts essential; comparison with existing featured articles on a similar topic helpful
B	content and style issues need to be resolved; check for supporting materials and compliance with style guidelines
C	significant gaps in content need to be addressed, cleanup issues resolved
Start	substantial improvements in content and organization are crucial; prioritize provision of references to reliable sources
Stub	lacking meaningful content; any editing or additional material helpful
FL	no further additions necessary unless new information becomes available
List	should contain appropriately named and organized live links to articles

1.2 Automatic Quality Assessment

The subsequently presented studies utilize Wikipedia's manual ratings as a foundation for the development of indicators and procedures to automatically assess an article's information quality. The authors have been addressed directly when there was need for clarification and to further supply information where necessary.

Stvilia et al. [13] use a sample of discussion pages in order to identify ten information quality problem types based on qualitative and quantitative characterizations. From three successive Wikipedia dumps they randomly extract 1 000 articles, which are further reduced to 841 by excluding stubs, redirects and deletions. Only 128 have non-empty discussion pages and contain more than 100 characters. In addition, 236 featured articles from one of the dumps build the foundation of a separate set, of which 235 articles' discussion pages meet the length criteria. 30 discussion pages from each set are selected and analyzed using the technique of content analysis. It is suggested that the process to improve the quality of an article is strongly connected to the data itself [14].

Blumenstock [15] takes a binary classification approach to distinguish featured from non-featured articles using a simple word count metric. They extract all articles from an English Wikipedia dump, of which they strip the markup, remove specialized files and articles containing less then fifty words. The featured

to non-featured ratio is approximately one to six. For the classification task they choose a threshold of 2 000 words, achieving 96.31% accuracy. However, the length of an article correlates with its quality iff the assumption holds that a featured status directly implies quality.

In a more comprehensive technical report Blumenstock [16] uses classification schemes implemented in R and WEKA[9] with approximately 100 extracted features per article. These schemes range from simple threshold functions to regression, random forest and multi-layer perceptron, to name a few. 5.6 million articles from a 2007 Wikipedia snapshot are used for evaluation. After removing articles with fewer than 50 words, templates, images, lists, metadata and Wikipedia-related markup and formatting, 1 554 featured and ca. a million non-featured articles remain. A random selection of the latter results in a corpus of 9 513 articles for training and testing.

Lipka and Stein [17] employ a technique comparable to word counts but yielding a higher discriminability. For the binary classification task they present the application of two learning algorithms, namely linear support vector machines (SVM) and Naïve Bayes (NB), to binarized character trigram vectors. The Biology and History domains of the English Wikipedia are the basis for two separate corpora providing 360 and 400 articles, half featured, half non-featured respectively. Three experiments are conducted, one using tenfold cross-validation [18], another one applying a classifier on a different domain than it was previously trained on and yet another, utilizing former classifier for the identification performance on three sets containing articles of different lengths.

For the task of automatically identifying an article's information quality flaw Anderka et al. [19] thoroughly analyzed specific *cleanup template messages* provided by the Wikipedia community. By employing various metrics used in information quality assessment in combination with new predictors a document model is developed, in which a document's feature vector represents a collection of quantifiable characteristics. In conjunction with a dedicated one-class learning approach and the assumption that featured articles contain no flaws, i.e., are flawless, the objective is to decide whether or not an article contains a certain flaw. They evaluate the effectiveness of their method using biased sample selection as well as the classification as a function of the flaw distribution.

Hu et al. [20] propose article quality measurement models based on the contribution of editors to provide a quality ranking. Taking into account how much

[9] http://www.cs.waikato.ac.nz/ml/weka/

an editor contributes by either authoring or editing an article they derive a final model called ProbReview. Only words within proximity of the edited part are considered when calculating the so-called reviewership. Several decaying schemes, i.e., functions of word distance are compared to a naïve metric, word count. 242 Wikipedia articles from the country domain are chosen for experimentation. The *Normalized Discounted Cumulative Gain at top k* (NDCG@k) metric and Spearman's rank correlation coefficient are used to evaluate the accuracy of a ranking and agreement between two rankings respectively.

Dalip et al. [21] study new, as well as existing quality indicators and use machine learning methods to combine numerous such indicators into one single assessment judgement. An article is represented as a vector of its features. They apply a regression method to find the best combination of the features to predict the quality value for any given article. For this to work, only the most discriminative features are chosen. They collect 874 articles according to the article quality grading scheme. The impact of the chosen indicators is evaluated using the information gain measure (infogain, for short) [18], whereas the classification performance is measured using the mean squared error (MSE). The ranking comparison metric NDCG@k is used to measure how close the predicted quality ranking of articles is to their true quality ranking. Dalip et al. [12] repeat the experiments with a greater number of collections and larger datasets in a new study, which yields similar results.

Others, such as Lih [8] study the trustworthiness of Wikipedia articles. Using features based on the edit history they evaluate the quality of articles after they were referenced in the press. While providing an insight into causes for quality improvement this research is not directly concerned with measuring article quality based on the features at hand. However, just like Wilkinson and Huberman [10], who study the difference in number of edits and editors, as well as the intensity of cooperative behavior regarding featured and non-featured articles, some useful benchmarks are presented, which may yield reasonable thresholds.

We implement only the most promising features and representations of the preceding studies, which are detailed in Chapter 3 and 4 respectively. A complete list of all the features employed in the various models and explored in the relevant literature can be found in Appendix A.1.

2 Retrieving Articles

There is an abundance of available methods to process Wikipedia's raw and metadata, most of which can be utilized through numerous ways online, using Web browsers or backups. We explore the data management and organization within the encyclopedia in Section 2.1 and conclude with a detailed explanation of the information extraction procedures in Section 2.2.

2.1 Data Organization in Wikipedia

Until January 2002, Wikipedia relied upon the already existing UseModWiki, an engine written in Perl, which stores all pages in individual text files with no history of any changes made. Articles were named using CamelCase to support automatic page linking before the double square brackets were introduced in a later patch.

Magnus Manske initiated the development of a dedicated Wikipedia engine called "PHP script" to remedy performance issues and other limitations imposed by generic wiki engines. It was written in PHP and uses a MySQL database for data storage. The application later served as the basis for the next and final iteration "Phase III". Many features were introduced that are still in use today, such as namespaces, skins and special pages. Persistent performance issues (due to increased traffic) and resource intensive features prompted another rewrite, which was done by Lee Daniel Crocker in 2002. The software has been continuously improved upon until June 2003, when Jimmy Wales created the Wikimedia Foundation. Two months later, it was officially named MediaWiki. Today, all of the foundations projects are powered by the free Web-based wiki software.

Table 2.1: Wikipedia basic namespaces. For each basic namespace exists a corresponding talk namespace, designated by adding "talk" to the default prefix.

Namespace	Content
Main	encyclopedia articles, lists, disambiguation pages and redirects
Project	pages connected with the Wikipedia project itself: information, policy, essays, processes, discussion, etc.
Portal	reader-oriented portals enabling subject specific grouping
User	public user pages for personal use
File	file description pages for image, video and audio files
MediaWiki	protected interface texts
Template	templates to be transcluded or substituted onto other pages
Category	pages displaying list of pages and subcategories
Book	Wikipedia books, collections of articles about one theme
Help	help pages for correct Wikipedia usage

2.1.1 Database Layout

As mentioned above, Wikipedia's current engine, MediaWiki, employs a MySQL database to store all the information. A complete overview of MediaWiki's MySQL database schema can be found at the Wikimedia website[1].

The layout comprises 53 tables, grouped by topic. Figure 2.1 identifies the tables used in this work. One of the most important features remains the organization in different namespaces, captured in the page table. Every one of the 22 (10 basic, 10 corresponding talk, 2 virtual) namespaces delineates a special subset of Wikipedia pages. The title indicates which namespace a page belongs to. For instance, the "User:" prefix assigns pages for personal use, which can, however, still be viewed and modified by others. Table 2.1 includes a complete list of all basic namespaces and their respective content. The main namespace, article namespace or mainspace contains all encyclopedia articles and is used without a prefix. For every other namespace the prefix is equal to its name, with the exception of the project namespace, the prefix of which can be either "Wikipedia:" or "WP:". Further information regarding Wikipedia's namespaces can be found on the internal project website[2].

[1] http://svn.wikimedia.org/viewvc/mediawiki/trunk/phase3/maintenance/tables.sql
[2] http://en.wikipedia.org/wiki/Wikipedia:Namespace

categorylinks	
* cl_from	INT(10) UNSIGNED
* cl_to	VARBINARY(255)
cl_sortkey	VARBINARY(230)
cl_timestamp	TIMESTAMP
cl_sortkey_prefix	VARBINARY(255)
cl_collation	VARBINARY(32)
cl_type	ENUM(...)

externallinks	
el_from	INT(8) UNSIGNED
el_to	BLOB
el_index	BLOB

page	
* page_id	INT(8) UNSIGNED
page_namespace	INT(11)
page_title	VARBINARY(255)
page_restrictions	TINYBLOB
page_counter	BIGINT(20) UNSIGNED
page_is_redirect	TINYINT(1) UNSIGNED
page_is_new	TINYINT(1) UNSIGNED
page_random	DOUBLE UNSIGNED
page_touched	BINARY(14)
page_latest	INT(8) UNSIGNED
page_len	INT(8) UNSIGNED

pagelinks	
* pl_from	INT(8) UNSIGNED
* pl_namespace	INT(11)
* pl_title	VARCHAR(255)

revision	
* rev_id	INT(8) UNSIGNED
* rev_page	INT(8) UNSIGNED
rev_text_id	INT(8) UNSIGNED
rev_comment	TINYBLOB
rev_user	INT(5) UNSIGNED
rev_user_text	VARBINARY(255)
rev_timestamp	BINARY(14)
rev_minor_edit	TINYINT(1) UNSIGNED
rev_deleted	TINYINT(1) UNSIGNED
rev_len	INT(8) UNSIGNED
rev_parent_id	INT(8) UNSIGNED
rev_shal	VARBINARY(32)

templatelinks	
* tl_from	INT(8) UNSIGNED
* tl_namespace	INT(11)
* tl_title	VARBINARY(255)

user	
* user_id	INT(10) UNSIGNED
user_name	VARBINARY(255)
user_registration	BINARY(14)

user_groups	
* ug_user	INT(5) UNSIGNED
* ug_group	VARBINARY(16)

Figure 2.1: Details of the Wikipedia database tables used in this work. Primary keys are marked with an asterisk (∗).

2.2 Exporting Mechanisms in Wikipedia

The Wikimedia Foundation maintains several projects, most of which employ the MediaWiki software. The Meta project, as such provides a forum for discussion and international coordination concerning all Wikimedia projects. It contains detailed information concerning the numerous access methods to the databases in use by the engine. Pages can be exported via direct HTTP requests in a special XML format, an example of which is specified in Listing 2.1. However, this kind of data mining consumes a lot of excess bandwidth and takes a long time to complete if a large amount of pages is required. A backup script periodically dumps all wiki pages and their metadata into separate XML files. These are obtainable through publicly hosted mirrors. Other mechanism, such as the Python Wikipedia Robot Framework and an OAI-PMH-interface exist, but are not available for everyone.

Listing 2.1: A simplified and condensed example of the exported page format. Three dots (...) indicate intentionally removed irrelevant text or metadata.

```
<mediawiki ... xml:lang="en">
  <siteinfo>
    ...
  </siteinfo>
  <page>
    <title>Information quality</title>
    <ns>0</ns>
    <id>1752647</id>
    <revision>
      <id>498526106</id>
      <parentid>490496691</parentid>
      <timestamp>2012-06-20T17:24:16Z</timestamp>
      <contributor>
        <username>Huzaifa1990</username>
        <id>16149032</id>
      </contributor>
      <minor/>
      <sha1>5x8jf5bbb4co519rqhf4396csaqzgi4</sha1>
      <text ... >'''Information quality''' (IQ) is a term to
          describe the quality of the content of [[information
          systems]]. It is often pragmatically defined as: &quot;
          The fitness for use of the information provided.&quot;</
          text>
    </revision>
  </page>
</mediawiki>
```

2.2.1 Index.php

Every MediaWiki site utilizes the *Index.php* script as the main entry point for HTTP requests, e.g., using a Web browser. Requests for the English Wikipedia are submitted via the following base URL:

```
http://en.wikipedia.org/w/index.php
```

It receives arguments as GET parameters, although some are passed as POST data. Every request contains a specific action, which, if not specified otherwise, defaults to "view", serving a page's normal content. Other actions, such as "delete" and "history" return the requested page in the corresponding form for deletion confirmation and history view, respectively.

This method does not efficiently scale in terms of data mining. The server instantiates a MediaWiki object for each request, followed by a title object depending on the action paramater. The objects are initialized after performing two additional checks before the final HTML markup is generated. Spidering pages in such a manner is considered bad practice and is generally discouraged.

2.2.2 Api.php

The MediaWiki API (*Api.php*), in contrast to the method above, provides high-level access to the database. Requests for the English Wikipedia are submitted via the following base URL:

```
http://en.wikipedia.org/w/api.php
```

The script supports a variety of output formats, including JSON, WDDX, XML, YAML, and serialized PHP. It was designed to be used in combination with server-side applications, such as bots, or thin web-based JavaScript applications, which log in to a wiki and alter its contents directly. It receives arguments as GET and POST parameters, similar to the Index.php method. A token has to be acquired by submitting a query action before any data can be modified. Each query submodule has its own namespace, which facilitate the generation of appropriate error codes upon failure. Privileged user or application accounts receive higher per-request limits to reduce the total time spent transforming the data. The API is slightly more efficient than the previous method when utilized for data mining.

Table 2.2: A list of the most important parameters for the Special:Export facility. Options are further restricted by MediaWiki specific configuration variables, e.g., `$wgExportMaxHistory` and `$wgExportAllowListContributors`.

Paramter	Description
addcat	include pages in categories as specified by catname
addns	include pages in namespaces as specified by nsindex
history	include the full history
limit	cumulative maximum number of revisions
listauthors	include a list of all contributors for each page
pages	page titles separated by linefeed characters
templates	include templates

2.2.3 Special:Export

The *Special:Export* tool constitutes another HTTP-based technique. Requests for the English Wikipedia are submitted via the following URL:

```
http://en.wikipedia.org/wiki/Special:Export
```

This method imposes fewer constraints when compared to the standard Web front-end and API. Articles of whole namespaces can be retrieved using the "AllPages" extension. It is also possible to request parts of or the entire revision history of the specified pages. Titles of desired articles can either be inserted into the above form, or they can be specified directly via GET parameters, the most important of which are featured in Table 2.2.

The exported pages comply with the above-mentioned XML format in Listing 2.1. The following URL returns the XML for the Article "Information quality", including every revision and a list of all contributors:

```
http://en.wikipedia.org/w/index.php?title=Special:
Export&pages=Information_quality&history&action=submit
```

2.2.4 Toolserver

The Wikimedia Toolserver represents an independent method with direct access to replicated databases of Wikipedia's raw and metadata. It can be accessed via the following URL:

https://wiki.toolserver.org/view/Main_Page

It is operated by the registered voluntary association Wikimedia Deutschland e. V. and provides Unix hosting for a multitude of tools administered by Wikimedia contributors. The project maintains 13 servers, organized in three clusters that contain a replica of Wikipedia and all its languages with different degrees of synchronization delay. The service is generally not available to the public; accounts on the Toolserver are provided for developers on a six months basis. The two types of user servers include Web based and Unix login servers, which are used to employ server-side applications, such as bots that manipulate data directly. This methodology promotes a platform optimized for application-oriented development, rather than large scale private retrieval, which is also reflected in the terms of use[3]:

> Tools may not serve significant portions of wiki page text to clients. "Significant" means distributing actual page content; [...]

2.2.5 Database Dump

The Wikimedia Foundation offers a complete copy of all its wikis, made accessible via the following URL:

http://dumps.wikimedia.org

A script periodically dumps all wiki pages into separate XML files, while other raw database tables in SQL form are made available directly. Wikimedia provides these monthly and bimonthly public backups for archival purposes, offline use and academic research. The process of backing up the largest of the wikis, enwiki, takes up to nine days to complete. The scripts parse the data sequentially, thus creating temporal inconsistencies when recording the state of the whole wiki; articles that are processed at a later stage may contain more current revisions. The backup comprises the most important content, such as pages, link tables, image metadata, and miscellaneous information, e.g., page properties and site statistics. The main page matter, the format of which complies with the XML output exemplified in Listing 2.1, is divided into three separate sets:

pages-articles.xml Includes the most current revision of all articles and templates, however, excludes pages from the user and discussion namespaces.

[3] https://wiki.toolserver.org/view/Rules

pages-meta-current.xml Includes the most current revision of all pages.

pages-meta-history.xml Includes every revision of all pages.

The smaller stub-articles.xml, stub-meta-current.xml and stub-meta-history.xml files contain the respective header information only.

Dump Preprocessing

The above method is best suited for the academic research in hand. We utilize a Hadoop cluster, which constitutes 44 computing nodes to store the database backup from January 2012. The pages-articles.xml contains 3 865 587 entries, of which 2 025 728 remain after removing non-article pages, articles with fewer than one interwiki link, disambiguation and empty pages, as well as stubs and lists. We employ a modified version of the Wikipedia Extractor[4] script (developed at the University of Pisa) to convert the Wikipedia markup to plaintext.

The dump preprocessing is illustrated in the activity diagram, which is depicted in Figure 2.2. With the additional SQL tables enumerated in Figure 2.1, the markup and plaintext form the basis for the feature computation, which is the subject of the next chapter.

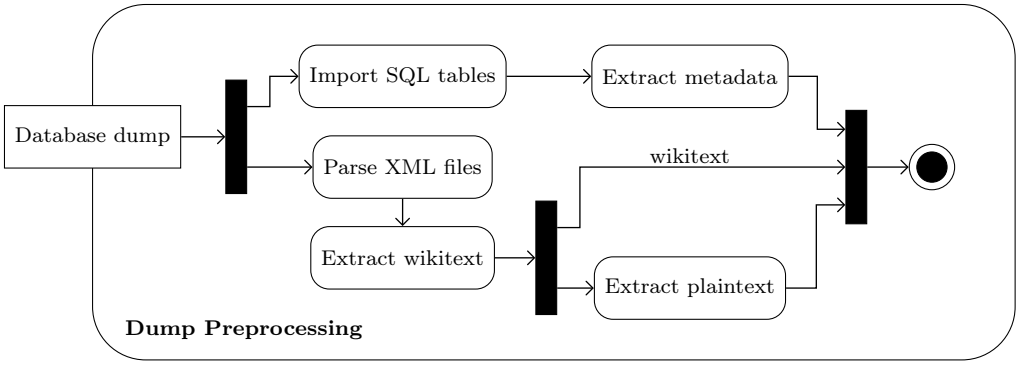

Figure 2.2: Schematic illustration of the dump preprocessing phase.

[4] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

3 Representing Articles

In this chapter we are going to examine the implementation details for the features used in the quality measurement procedures. Most of the related work presented in Chapter 1 apply the following model, as formulated in depth by Dalip et al. [12]. Let $A = \{a_1, a_2, \ldots, a_n\}$ be a set of n articles, each represented by m features $F = \{f_1, f_2, \ldots, f_m\}$. A vector representation for each article a_i in A is defined as $a_i = (v_1, v_2, \ldots, v_m)$, where v_j is the value of feature f_j. A feature generally describes some quality indicator associated with an article. A few differ slightly from one another, e.g., counts divided by the number of characters instead of words or ratios instead of a pure count. More commonly no explanation is supplied at all. In such cases we choose the solution which provides the best results—often ratios prove to be the better choice over simple counts.

All article features are organized along the four dimensions content, structure, network, and history, as proposed by Anderka et al. [19]. Because of the sheer amount of different quality indicators and expensive computation of network and history algorithms, not every metric is used in the final evaluation. We refer to the results in [12]:

> Through experiments, we show that the most important quality indicators are the easiest ones to extract, namely, textual features related to length, structure and style. We were also able to determine which indicators did not contribute significantly to the quality assessment. These were, coincidentally, the most complex features, such as those based on link analysis.

A complete overview of all the features described below, along with a reference to their origin, can be found in Appendix A.1. Features not used in this work due to time constrains or insufficient information are marked with a small cross (×) in the aforementioned appendix. Some of these were originally not intended to be used for quality assessment.

3.1 Content Features

The NLTK[1] Python package is used to extract most of the components used in the features that follow. We use the Punkt sentence tokenizer from the aforementioned package to separate the sentences. Words are filtered using the corresponding Punkt word tokenizer before removing digits and punctuation. These are then tagged using a previously trained part of speech tagger based on the Brown corpus[2] to identify lexical categories.

Character count Number of characters in the plaintext, without spaces.

Word count Number of words in the plaintext.

Sentence count Number of sentences in the plaintext.

Word length Average word length in characters as defined in Equation 3.1.

$$wl = \frac{characterCount}{wordCount} \tag{3.1}$$

Sentence length Average sentence length in words as defined in Equation 3.2.

$$sl = \frac{wordCount}{sentenceCount} \tag{3.2}$$

Syllable count Number of syllables in the plaintext. Greg Fast's Perl module Lingua::EN::Syllable[3] is used to estimate the count for each word.

Word syllables Average number of syllables per word.

One-syllable word count Number of one-syllable words.

One-syllable word rate Percentage of one-syllable words as defined in Equation 3.3.

$$syr = 100 \cdot \frac{oneSyllableWordCount}{wordCount} \tag{3.3}$$

[1] http://nltk.org/api/nltk.html
[2] http://www.comp.leeds.ac.uk/ccalas/tagsets/brown.html
[3] http://search.cpan.org/~gregfast/Lingua-EN-Syllable-0.251/Syllable.pm

Paragraph count Number of paragraphs. These are typically separated by two newline in the raw dump data.

Paragraph length Average paragraph length in sentences.

Longest sentence length Number of words in the longest sentence.

Shortest sentence length Number of words in the shortest sentence.

Long sentence rate Percentage of long sentences. A long sentence is defined as containing at least 30 words.

Short sentence rate Percentage of short sentences. A short sentence is defined as containing at most 15 words.

Passive sentence rate Percentage of passive voice sentences.

Sentence beginning rate Percentage of sentences beginning with either a pronoun, interrogative pronoun, article, conjunction, subordinate conjunction or preposition.

Question rate Percentage of questions.

Auxiliary verb rate Percentage of auxiliary verbs.

Conjunction rate Percentage of conjunctions.

Nominalization rate Percentage of nominalizations. A nominalization is the product of adding a derivational suffix to a part of speech, usually a verb, adjective or adverb. In this work we use the same technique applied in the GNU Style[4] software – a word is a nominalization if the suffix is equal to either "tion", "ment", "ence" or "ance".

Preposition rate Percentage of prepositions.

Pronoun rate Percentage of pronouns.

"To be" verb rate Percentage of "to be" verbs.

Special word rate Percentage of personal, interrogative, relative, indefinite and demonstrative pronouns such as weasel, peacock, doubt and editorializing words, as well as idioms, aphorisms and proverbs.

[4] http://www.gnu.org/software/diction/

Information-to-noise ratio Percentage of unique words. We identify the bag of words (bowSize) after discarding all stop words and stemming the remainder. The ratio is computed as defined in Equation 3.4.

$$itn = \frac{bowSize}{wordCount} \tag{3.4}$$

Automated Readability Index Readability formula (3.5) developed by Smith and Senter [22]. They determine the US grade level needed to comprehend an English text. The formula relies on the average of characters per word and the average of words per sentence.

$$ari = 4.71 \cdot \frac{characterCount}{wordCount} + 0.5 \cdot \frac{wordCount}{sentenceCount} - 21.43 \tag{3.5}$$

Coleman-Liau Index Readability formula (3.6) developed by Coleman and Liau [23] to approximate the usability of a text, utilizing the average number of letters and sentences per 100 words.

$$cli = 5.89 \cdot \frac{characterCount}{wordCount} + 30 \cdot \frac{sentenceCount}{wordCount} - 15.8 \tag{3.6}$$

Flesch reading ease Readability formula (3.7) developed by Flesch [24] to indicate comprehension difficulty when reading a passage of contemporary academic English. They use the average number of syllables per word and average sentence length to compute a value between 0 and 100, where 0 indicates a text hard to read.

$$fres = 206.835 - 84.6 \cdot \frac{syllableCount}{wordCount} - 1.015 \cdot \frac{wordCount}{sentenceCount} \tag{3.7}$$

Flesch-Kincaid Readability formula (3.8) based on the above metric, developed by Kincaid et al. [25] to indicate comprehension difficulty when reading a passage of contemporary academic English. They use the same measurements translating the score to the respective US grade level.

$$fkgl = 0.39 \cdot \frac{wordCount}{sentenceCount} + \frac{syllableCount}{wordCount} - 15.59 \tag{3.8}$$

FORCAST readability Readability formula developed by Caylor and Sticht [26], intended for short texts containing 150 words. The number of one-syllable words and the total number of words serve as core measures. In order to make use of this metric we adapt the original formula for arbitrarily sized texts yielding Equation 3.9.

$$fgl = 20 - \frac{oneSyllableWordCount}{10 \cdot \frac{wordCount}{150}} \tag{3.9}$$

Gunning Fog Index Readability formula (3.10) developed by Gunning [27]. They measure the readability of English writing by utilizing the average sentence length and percentage of complex words (defined as words containing more than three syllables).

$$gfi = 0.4 \cdot \left(\frac{wordCount}{sentenceCount} + 100 \cdot \frac{complexWordCount}{wordCount} \right) \tag{3.10}$$

Läsbarhetsindex Readability formula (3.11) developed by Björnsson [28] to assess text difficulty in foreign languages. Here, complex words are defined as words with seven or more characters rather than as the count of syllables. Greater scores suggest a higher level of text difficulty, although they are interpreted differently depending on the language used in the text.

$$lix = \frac{wordCount}{sentenceCount} + 100 \cdot \frac{complexWordCount}{wordCount} \tag{3.11}$$

SMOG Grading Readability formula (3.12) developed by McLaughlin [29] to estimate the necessary years of education to understand a specific piece of text. Since its creation in 1969, the formula has continuously been improved upon[5].

[5] http://webpages.charter.net/ghal/SMOG.htm

$$sgi = \sqrt{30 \cdot \frac{complexWordCount}{sentenceCount}} + 3 \tag{3.12}$$

Trigrams Feature developed by Lipka and Stein [17]. A trigram vector of characters is represented as a mapping from substrings of three tokens to their respective frequencies. The binarized vector maps to the occurrence of a trigram instead of its frequency. In combination with a linear SVM their experiments yield the most promising results.

3.2 Structure Features

Most of the structural features are count-based or concerned with their respective ratio to text length and provide information about the organization of an article. Text length is computed on the plaintext, whereas syntactic elements, such as the link or section count, are derived from the unfiltered markup.

Section count Number of sections. As defined by the MediaWiki markup, section headings are introduced by multiple equal signs (=). We use the following regular expression to match sections:

```
^={2}([^=]+?)={2}$
```
(3.13)

Subsection count Number of subsections. The count of equal signs in the regular expression (3.13) is increased to three in order to match subsections.

Heading count Number of sections, subsections and subsubsections.

Section nesting Average number of subsections per section.

Subsection nesting Average number of subsubsections per subsection.

Section distribution Average heading length in words. For each section, subsection and subsubsection the average length (e.g., for sections in Equation 3.14) in words is used to derive an overall average.

$$asl = \frac{\sum_{i=1}^{n} sectionLength_i}{n} \tag{3.14}$$

Longest section length Number of words in the longest section.

Shortest section length Number of words in the shortest section.

Trivia section count Number of trivia sections. The section names include "facts", "miscellanea", "other facts", "other information" and "trivia".

Reference section count Number of reference sections. An exhaustive list of references section names can be found in Appendix A.2.

Lead length Number of words in the lead section. A lead section is defined as the text before the first heading. Without a heading there is no lead section.

Lead rate Percentage of words in the lead section.

Link rate Percentage of links. Every occurrence of a link (introduced with two open square brackets) in the unfiltered article text is considered when computing the ratio of link count to word count in the plaintext. This feature is not as representative as the out-link rate.

Reference count Number of all references using the `<ref>...</ref>` syntax, including citations and footnotes.

Reference section rate Ratio of reference count to the accumulated section, subsection and subsubsection count.

Reference word rate Ratio of reference count to word count.

Image count Number of images.

Image rate Ratio of image count to section count.

Table count Number of tables.

File count Number of linked files.

Category count Number of Wikipedia categories an article belongs to, derived from the category link.

Template count Number of unique Wikipedia templates.

List rate Ratio of words in lists to word count. List items are defined as lines starting with an asterisk, a pound character or a semicolon.

3.3 Network Features

This section addresses quality indicators derived from internal and external page links extracted from the Wikipedia database tables as seen in Figure 2.1.

Internal link count Number of outgoing internal links, querying the `pagelinks` table as specified in Listing 3.1. The `internal.article` table contains all page IDs and their respective titles from the Wikipedia's article namespace. An internal link is defined as a hyperlink pointing to another article within the Wikipedia domain.

Listing 3.1: Internal link count SQL query.

```
SELECT pl_from,
       COUNT(DISTINCT pl_title)
FROM   pagelinks
       JOIN internal.article
         ON pl_from = page_id
WHERE  pl_namespace = 0
GROUP  BY pl_from
```

Broken internal link count Number of broken internal links, which is defined as an outgoing link to an article that does not exist. The left join operation (Listing 3.2) produces a list of links which are not part of the aforementioned article namespace.

Listing 3.2: Broken internal link count SQL query.

```
SELECT p.pl_from,
       COUNT(DISTINCT p.pl_title)
FROM   (SELECT pl_from,
               pl_title
        FROM   pagelinks
        WHERE  pl_namespace = 0
        GROUP  BY pl_title) AS p
       LEFT JOIN (SELECT page_id,
                         page_title
                  FROM   internal.article
                  ORDER  BY page_title) AS a
              ON p.pl_title = a.page_title
WHERE  a.page_id IS NULL
GROUP  BY p.pl_from
```

External link count Number of outgoing external links, querying the `externallinks` table. An external link is defined as a hyperlink pointing to a document outside the Wikipedia domain. The query is similar to the one specified in Listing 3.1.

External links per section Number of external links per section.

Language link count Number of outgoing language links, querying the `langlinks` table. A language link is defined as a hyperlink pointing to the same article in another language. The query is similar to the one specified in Listing 3.1.

In-link count Number of incoming internal links, querying the `pagelinks` and page tables. The query is similar to the one specified in Listing 3.1.

Reciprocity Ratio between the number of articles referencing an article and those referenced by the same. We assume this to be equivalent with the ratio of in-link count to internal link count.

Clustering coefficient Ratio between the number of k-nearest-neighbors and the number of possible edges between a node and its k-nearest-neighbors. In Equation 3.15, $edgesCount(k)$ is defined as the number of nodes which have a path to n whose length is, at most, k edges, whereas $maxEdges(k)$ describes the latter.

$$cc = \frac{edgesCount(k)}{maxEdges(k)} \tag{3.15}$$

Reference measures Assortativity of a node defined as ratios of in-link and internal link count to the average in-link and internal link count of the node graph neighbours. This metric has been developed to detect spam in websites and online video systems.

PageRank PageRank value of an article computed according to Brin and Page [30]. It is suggested in [12] that the importance of an article is proportional to the importance and quantity of articles that point to it. We build a link graph of all articles and compute the page rank.

3.4 History Features

The following features are extracted from the edit history of each article. This is accomplished by parsing the pages meta history dump. When no time span is supplied with a feature description the period from creation (first revision) to now (date of the snapshot) applies.

Age Number of days since an article's creation, extracted from the current edit's timestamp.

Age per edit Ratio between age and number of edits on the last 30 revisions. We assume this to be the average time past between each of the revisions.

Currency Number of days between the last edit and now (date of the snapshot).

Edit count Number of edits.

Edit rate Percentage of edits per day and per user.

Edit distribution Standard deviation of edits per user.

Edit currency rate Percentage of edits made in the last three months.

Editor count Number of distinct editors.

Editor role Number of administrative, anonymous and registered users.

Editor rate Percentage of edits made by the top 5% of most active editors and by users who edited the article less than four times.

Connectivity Number of articles with common editors. Two articles share an editor when at least one of their revisions was made by the same user.

Discussion count Number of edits of an article's discussion page.

Revert count Number of reversions. We compute an MD5 hash of the text content of each revision. Every duplicate indicates a reversion of an article.

Revert time Average revert time in minutes. The revert time is defined as the time between an edit and its reversion.

Modified lines rate Number of modified lines when comparing the current version of an article to the one three-months old.

Review Review, BasicReview and ProbReview are features proposed by Hu et al. [20] to measure the quality of articles based on the quality of their editors. Each article is considered a bag of words, such that $a_i = w_{ik}$. The quality Q_i of an article i sums up all word qualities q_{ik} as depicted in Equation 3.16.

$$Q_i = \sum_k q_{ik} \tag{3.16}$$

The final and most accurate model ProbReview is defined in Equations 3.17 and 3.18 as follows.

$$q_{ik} = \sum_j f(w_{ik}, u_j) \cdot A_j \tag{3.17}$$

$$A_j = \sum_{i,k} f(w_{ik}, u_j) \cdot q_{ik} \tag{3.18}$$

where,

$$f(w_{ik}, u_j) = \begin{cases} 1 & \text{if } w_{ik} \overset{A}{\leftarrow} u_j \\ Prob\left(w_{ik} \overset{R}{\leftarrow} u_j\right) & \text{otherwise} \end{cases} \tag{3.19}$$

A_j, the authority of each user u_j, is obtained by the summation of the quality of each word q_{ik} multiplied by the review probability of a word w_{ik}. Latter is based on the proximity to the closest authored word w_{il} by the user. The relationship between a user and a word are denoted by:

- $w_{ik} \overset{A}{\leftarrow} u_j$, word w_{ik} is authored by user u_j
- $w_{ik} \overset{R}{\leftarrow} u_j$, word w_{ik} is reviewed by user u_j

The function $Prob(w_{ik} \overset{R}{\leftarrow} u_j)$ is defined to return 0 when the user u_j has never updated the content of article a_i, or when the word w_{ik} has never appeared in the user's edit(s) of the article. In all other cases the function returns the review probability of a word w_{ik} as expressed in the monotonically decaying function (3.20).

$$S(d_{kl}) = \frac{1}{\sqrt{\max(|d_{kl}| - \alpha, 0) + 1}}, \tag{3.20}$$

where $\alpha = 7$ and d_{kl} is defined as the distance between w_{ik} and w_{il}. The choice of α is based on empirical studies.

4 Classifying Articles

We present the construction and evaluation of the experiments to classify featured and non-featured articles. To ensure the reproducibility, the previous chapter describes every article feature found in the relevant literature in detail. We consider only those studies that are based on the proposed vector space representation. Furthermore, we provide an overview of the classification methods utilized in each study in Section 4.1. Before discussing the results of the classification experiments in Section 4.2, we specify the evaluation measures which are used to compute their effectiveness.

4.1 Classification Methodology

Seven of the works presented in Chapter 1 apply the article representation proposed in the previous chapter, only four of which supply sufficient information about the classification methods used to automatically assess the quality of Wikipedia articles. The following detailed overview summarizes the methods employed in each of the studies, along with the difficulties encountered during the respective replication. Table 4.1 contains the details of the article selections taken into account for evaluating the listed classifiers, which are also clarified in the summary below.

Table 4.1: Article distribution for datasets used to construct the classifiers found in the relevant literature, along with the classifiers applied in this work.

Model	**Articles**		**Selection Strategy**	**Classifier**
	featured	non-featured		
(1) Blumenstock	1 554	9 513	random	MLP
(2) Dalip et al.	549	2 745	random	SVM
(3) Lipka and Stein	380	380	domain-specific	SVM
(4) Stvilia et al.	236	834	random	C4.5

(1) Blumenstock

In contrast to complex quantitative methods, Blumenstock [15] proposes a single metric, the length (word count) of an article, as its sole representation. To evaluate the performance he chooses an unbalanced subset (ratio 1:6, featured : non-featured) of the English Wikipedia and multiple classification techniques, such as a logit model, a rule-based, k-nearest neighbor and random-forest classifier, as well as a multi-layer perceptron (MLP). The latter achieves the highest overall accuracy and thus serves as a baseline for the replication of the experiment. Similar to the author, we utilize the WEKA implementation of the algorithm and further tune the settings to include 1 000 epochs to train through and a validation set size of 30%. The classifier is additionally evaluated by tenfold cross-validation[1]. However, we suspect the performance to gradually degrade with the increase of the minimum word count.

(2) Dalip et al.

Dalip et al. [12] provide a detailed summary of many article features found in related work. They analyze the impact of each feature dimension in the quality assessment of online encyclopedia articles by applying the information gain measure[2]. The effectiveness of the proposed classification method is evaluated using MSE, which is defined in Equation 4.1.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} e^2, \tag{4.1}$$

where e is the error value and n is the number of articles. They apply a support vector regression (SVR), a radial basis function as the kernel type and tenfold cross-validation to estimate an article's quality class. We contacted the authors who supplied us with the exact page IDs and feature values of their conducted studies. The replicated experiment yields similar results when utilizing the article features we described in the previous chapter.

Even though we are able to compute the mean squared error in this manner, the actual accuracy with which an article is successfully classified

[1] The dataset is randomly split into ten parts, one of which serves as a test set, while the remaining parts act as the training set [18].

[2] A statistical measure that indicates the contribution of a given feature to discriminate the class to which any given article belongs [18].

is withheld. Therefore, we choose to apply the linear SVM implementation from WEKA's libSVM library instead. Additional settings include the normalization of the input data. Due to insufficient information it is unknown to which effect the information gain measure is utilized. Missing feature values need to be replaced by executing an attribute filter beforehand, since WEKA's implementation of the algorithm cannot handle these.

(3) Lipka and Stein

Articles are represented by writing-style-related (syntactic) features and their binarizations: character trigram vectors and part of speech trigram vectors. Lipka and Stein [17] apply two machine learning algorithms (SVM, NB) and evaluate the classifier by tenfold cross-validation within a single domain to minimize the influence of topical discrimination. The combination of a linear SVM with a binarized character trigram vector representation yields the highest identification performance of featured articles.

In order to replicate their study, we compute the character trigram vectors for each of the articles extracted during the dump preprocessing phase, a trivial task when leveraging the power of the Hadoop cluster. To retrieve domain-specific datasets, a recursive query returns all respective article page IDs, from which a sample is randomly selected. We continue with the construction of the character trigram vocabulary and pair the frequency of each trigram within an article with its position in the vocabulary to create a sparse arff file[3]. The latter is a necessary step to allow for an efficient classification, since a vector can easily exceed 50 000 features.

The results provided by Lipka and Stein [17] show that the binarized representations outperform the non-binarized. However, during the course of our experimentation we observed the opposite—the accuracy to which an article is correctly classified decreases by 10% on average when choosing a binarized representation. Based on this result, we decided on non-binarized vectors to represent the articles. Furthermore, the 500 most discriminative character trigrams, ranked by information gain, supply the features for the classification with WEKA's linear SVM algorithm. The input data is normalized similarly to the classification method described in (2) above.

[3] http://www.cs.waikato.ac.nz/ml/weka/arff.html

(4) Stvilia et al.

Based on variable groupings (identified through exploratory factor analysis) Stvilia et al. [31] propose 19 distinct quality indicators as a mean to represent articles. Employing the WEKA implementation of the Density Based Clustering algorithm they analyze the distribution based on the information quality measures. Yet another WEKA implementation, the C4.5 Decision Tree algorithm (J48) is utilized in combination with tenfold cross-validation to test the metrics.

We replicate the binary classification experiment by filtering the specified indicators from the whole set of extracted features.

(5) Fricke and Anderka

In addition to the studies presented above, we propose our own assessment model by taking all of the article features implemented in this work into account—except for the trigrams feature, which requires a representation on its own.

The experiment setup is equivalent to the one described in (2) above.

4.1.1 Evaluation Measures

In a classification setting the evaluation measures *Precision* and *Recall* are defined in terms of true positives (tp – the number of items correctly labeled as belonging to the positive class), true negatives (tn – the number of items correctly labeled as belonging to the negative class), false positives (fp – the number of items incorrectly labeled as belonging to the positive class, also known as type I errors) and false negatives (fn – the number of items incorrectly labeled as belonging to the negative class, also known as type II errors).

Precision and *Recall*, which are also referred to as specificity and sensitivity, are defined for the positive class in Equations 4.2 and 4.3, respectively.

$$Precision = \frac{tp}{tp + fp} \tag{4.2}$$

$$Recall = \frac{tp}{tp + fn} \tag{4.3}$$

Table 4.2: Confusion matrix example: 99 and 9 items are correctly labeled as featured (positive) and non-featured (negative) articles, while 11 and 1 items are incorrectly labeled as featured (positive) and non-featured (negative) articles, respectively.

		Actual class	
		featured	non-featured
Predicted class	featured	99	11
	non-featured	1	9

These are best observed in a confusion matrix, a layout that allows a quick visualization of the performance of supervised learning algorithms, an example of which is depicted in Table 4.2. Each column represents instances of the actual class, whereas each row represents instances of the predicted class. In regard to the example, this translates to 99 out of 100 correctly classified featured articles and 9 out of 20 correctly classified non-featured articles.

The F-Measure, also referred to as F_1 or balanced F-score, combines $Precision$ and $Recall$ to form the harmonic mean depicted in Equation 4.4. It is a special case of the general F_β, where $\beta = 1$, thus evenly weighting $Precision$ and $Recall$.

$$F\text{-Measure} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{4.4}$$

Unbalanced datasets may negatively influence the average values of the measures, given a high enough disparity between $Precision$ and $Recall$. Class distribution weighted averages alleviate this problem to a certain extent.

4.2 Experiment Conclusion

We conduct two kinds of experiments to evaluate the performance of the models described in the previous section. The first concerns itself with the replication of the original experiments as documented in the relevant literature. Most of the studies employ a randomized article selection strategy while not specifying the date of the snapshot or the revisions used for the assessment. Due to its editing nature, Wikipedia's articles are in constant flux, which results in

Table 4.3: Performance comparison of original (first row) and replicated (second row) experiments. ⊥ indicates information not provided in the literature. Model: (1) Blumenstock, (2) Dalip et al., (3) Lipka and Stein, (4) Stvilia et al.

Model	**Featured**	**Non-featured**	**Average**
	Precision/ *Recall*/ *F*-Measure	*Precision*/ *Recall*/ *F*-Measure	*F*-Measure
(1)	0.871 / 0.936 / 0.902	0.989 / 0.977 / 0.983	0.970
	0.781 / 0.877 / 0.826	0.980 / 0.960 / 0.970	0.949
(2)	⊥	⊥	⊥
	0.903 / 0.900 / 0.901	0.980 / 0.981 / 0.980	0.967
(3)	0.966 / 0.961 / 0.964	⊥	⊥
	0.949 / 0.939 / 0.944	0.940 / 0.950 / 0.945	0.944
(4)	0.900 / 0.920 / 0.910	0.980 / 0.970 / 0.975	0.957
	0.859 / 0.907 / 0.882	0.973 / 0.958 / 0.965	0.947

ever-evolving data. For this reason, the assembly of exact copies of the original datasets is impossible.

A snapshot of the English Wikipedia's database from January 2012 serves as the basis for all the conducted experiments. After preprocessing the backup dump (Section 2.2.5) we compute every implemented feature for each article by utilizing the extracted wiki- and plaintext, as well as the metadata contained in the imported MySQL tables. In combination these metrics provide the necessary features for the models (detailed in Table 4.1), for which the datasets are defined based on the given constrains. Finally, we convert the evaluation sets to a WEKA suitable file format and evaluate each classifier as described above.

Table 4.3 presents a performance comparison of the original and replicated experiments. Unfortunately not all of the previous work provide their results in form of *Precision*, *Recall* and *F*-Measure. However, the values for the measures yielded by the replication show how well each classifier performs on the same Wikipedia snapshot. For the studies which supply the data, we achieve a remarkably close match.

Except for (3) all models generally fail to classify featured articles with a similar efficiency when compared to non-featured. The cause lies within the choice of the article distribution. Given two classes, increasing the *Recall* (*Precision*) of the positive class increases the *Precision* (*Recall*) of the negative

class, respectively. The amount by which one measure influences the other is determined by the class distribution. This is the rationale behind the balanced distribution of featured and non-featured articles selected for the following series of experiments.

We define four new datasets corresponding to the minimum article length in words (0, 800, 1 600, 2 400) by randomly sampling 6 000 featured and non-featured articles over all domains. The selection strategy ensures a uniform distribution over the whole of Wikipedia's article space. After constructing the feature vectors for each model we evaluate the classifiers accordingly, the performance of which is captured in Table 4.4—visually supported by six plots in Figure 4.1. An additional model (5) is introduced to further explore the effectiveness of combining all features into a single assessment judgement.

Two significant observations can be made. The *Recall* of the featured class positively correlates with the *Precision* of the non-featured class and vice versa. This behavior is due to the balanced property of the evaluation sets and the defining characteristics of the measures. Equation 4.5 defines the *Recall* from the perspective of the negative class.

$$Recall_n = \frac{tn}{tn + fp} \tag{4.5}$$

Because of the false positives fp, Equation 4.2 and 4.5 are in a constant relation. Therefore, one measure can be explained by means of the other (4.6).

$$Precision \simeq Recall_n \tag{4.6}$$

The second observation concerns the slight, yet gradual degradation in performance of each model corresponding to the increase of the minimum word count. The classifiers are most effective when no threshold is applied and prove less so the larger the threshold becomes. By excluding short articles from the evaluation sets the *Recall* (retrieval) of non-featured articles decreases, validating the results found in Blumenstock [15]: the quality correlates directly with the length of an article. Furthermore, the outcome of our experiments shows that the more complex models significantly outperform the simple ones. It should be noted that (3) belongs to the complex category because of its expensive computational cost.

The classifier trained with our model (5) achieves the best overall performance, followed closely by (2), the selected features of which can technically be considered a subset of ours.

Table 4.4: Performance comparison grouped by the article length using balanced datasets with 6 000 featured and non-featured articles randomly sampled over all domains. The maximum F-measure values are marked in bold. Model: (1) Blumenstock, (2) Dalip et al., (3) Lipka and Stein, (4) Stvilia et al., (5) Fricke and Anderka.

Model	**Featured**	**Non-featured**	**Average**
	Precision / *Recall* / *F*-Measure	*Precision* / *Recall* / *F*-Measure	*F*-Measure
Word count > 0.			
(1)	0.925 / 0.958 / 0.941	0.957 / 0.922 / 0.939	0.940
(2)	0.955 / 0.974 / 0.964	0.973 / 0.954 / 0.963	0.964
(3)	0.951 / 0.967 / 0.959	0.966 / 0.950 / 0.958	0.958
(4)	0.925 / 0.974 / 0.940	0.972 / 0.921 / 0.946	0.947
(5)	0.958 / 0.982 / **0.970**	0.982 / 0.957 / **0.969**	**0.970**
Word count > 800.			
(1)	0.920 / 0.961 / 0.940	0.959 / 0.917 / 0.937	0.939
(2)	0.956 / 0.975 / 0.965	0.974 / 0.955 / 0.965	0.965
(3)	0.952 / 0.965 / 0.958	0.964 / 0.951 / 0.958	0.958
(4)	0.925 / 0.962 / 0.943	0.960 / 0.922 / 0.941	0.942
(5)	0.960 / 0.984 / **0.972**	0.983 / 0.959 / **0.971**	**0.971**
Word count > 1600.			
(1)	0.913 / 0.952 / 0.932	0.950 / 0.910 / 0.929	0.931
(2)	0.953 / 0.974 / 0.964	0.974 / 0.952 / 0.963	0.963
(3)	0.952 / 0.962 / 0.957	0.961 / 0.951 / 0.956	0.956
(4)	0.920 / 0.963 / 0.941	0.961 / 0.916 / 0.938	0.939
(5)	0.955 / 0.978 / **0.966**	0.978 / 0.954 / **0.966**	**0.966**
Word count > 2400.			
(1)	0.893 / 0.958 / 0.924	0.955 / 0.885 / 0.918	0.921
(2)	0.944 / 0.967 / 0.955	0.966 / 0.943 / 0.954	0.955
(3)	0.945 / 0.948 / 0.946	0.948 / 0.945 / 0.946	0.946
(4)	0.905 / 0.957 / 0.930	0.954 / 0.900 / 0.926	0.928
(5)	0.950 / 0.974 / **0.962**	0.974 / 0.949 / **0.961**	**0.961**

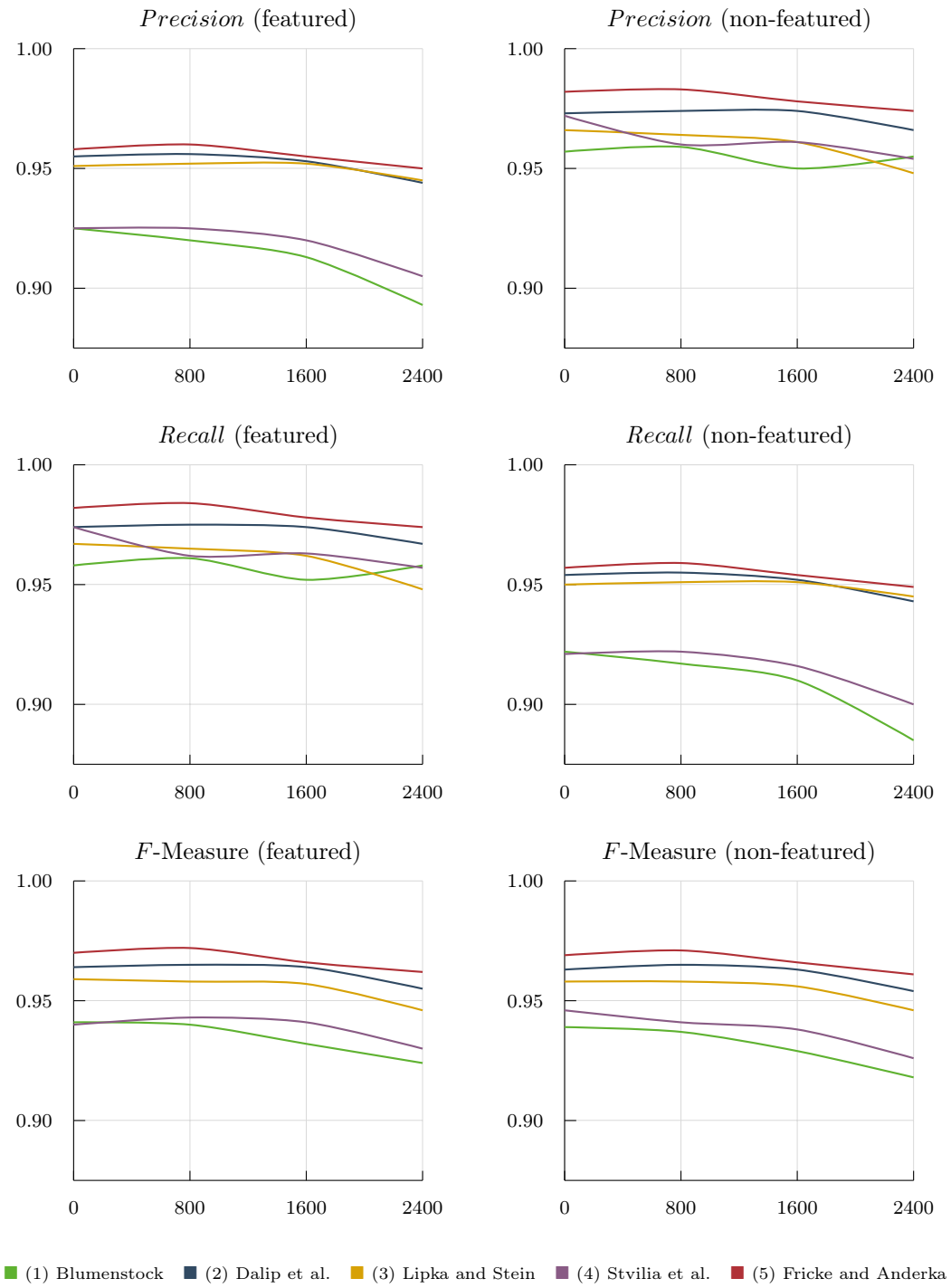

Figure 4.1: Evaluation measures (y-axis) over minimum word count (x-axis) corresponding to the performance comparison results found in Table 4.4.

5 Conclusion

The concept of featured articles is part of Wikipedia's quality grading scheme. In our study it served as the basis for automatically assessing information quality by means of classification, for which current models differ in the representation of articles. Furthermore, the classifiers are separately evaluated on a number of different datasets. We created a framework to consistently evaluate multiple models and introduced a new representation based on Dalip et al. [12], which has been the best performing model so far. Our model outperforms the latter and the baseline by 0.005 and 0.034, respectively, in terms of the F-Measure. Regarding the entire Wikipedia the number of articles correctly classified as featured and non-featured increases by approximately 30 000. The implementation of the framework utilizes Hadoop and scales with the number of articles, reducing the time needed from feature computation to classifier evaluation to a couple of hours.

As a major contribution we presented the most comprehensive collection of article features to date, almost all of which have been implemented and employed in the various models.

Future work could include further exploration of novel quality indicators and models for classification. Moreover, in combination with flaw detection algorithms the framework could be used to improve upon the quality of articles in poor condition.

A Appendix

A.1 Complete Feature List

Table A.1: A complete list of Wikipedia article features. Metrics not implemented in this work are marked with a small cross (×).

Feature	Reference
Derived from Content.	
Automated Readability Index	[12, 16]
Auxiliary verb rate	[12, 19]
Character count	[12, 16, 31]
Coleman-Liau Index	[12, 16]
Conjunction rate	[12]
FORCAST readability	[16]
Flesch reading ease	[12, 16, 31]
Flesch-Kincaid	[12, 16, 31]
Gunning Fog Index	[12, 16]
Information-to-noise ratio	[31]
Long sentence rate	[12]
Longest sentence length	[12, 19]
Läsbarhetsindex	[12]
Nominalization rate	[12]
One-syllable word count	[16]
One-syllable word rate	[19]
Paragraph count	[19]
Paragraph length	[12]
Passive sentence rate	[12]
Preposition rate	[12]
Pronoun rate	[12, 19]
Question rate	[12, 19]

continued on next page

Table A.1 (continued)

Feature	**Reference**
SMOG Grading	[16]
Sentence beginning rate	[12, 19]
Sentence count	[12, 16]
Sentence length	[19]
Short sentence rate	[12]
Shortest sentence length	[19]
Special word rate	[19]
Syllable count	[16]
Trigrams	[17]
Word count	[16, 17, 20]
Word length	[19]
Word syllables	[19]
"To be" verb rate	[12]
Derived from Structure.	
Category count	[16]
File count	[16]
Heading count	[16, 19]
Image count	[12, 16, 31]
Images per section	[12]
Lead length	[12, 19]
Lead rate	[19]
Link rate	[12, 19]
List rate	[19]
Longest section length	[12]
Reference count	[12, 16]
Reference section count	[19]
Reference section rate	[12]
Reference word rate	[12]
Section count	[12, 16]
Section distribution	[12]
Section nesting	[12]
Shortest section length	[12]
Subsection count	[12]
Subsection nesting	[19]
Table count	[16]

continued on next page

Table A.1 (continued)

Feature	**Reference**
Template count	[19]
Trivia section count	[19]
Derived from Network.	
Broken internal link count	[31]
× Clustering coefficient	[12]
External link count	[12, 16, 31]
External links per section	[12]
In-link count	[19]
Internal link count	[12, 16, 31]
Language link count	[12]
PageRank	[12]
Reciprocity	[12]
× Reference measures	[12]
Derived from History.	
Age	[12, 31]
Age per edit	[12]
× Connectivity	[31]
Currency	[31]
Discussion count	[10, 12]
Edit count	[8, 10, 12, 31]
Edit currency rate	[12]
× Edit distribution	[12]
× Edit rate	[10, 12]
Editor count	[8, 10, 31]
× Editor rate	[12]
× Editor role	[12, 31]
× Modified lines rate	[12]
Revert count	[31]
Revert time	[31]
× Review	[20]

A.2 Reference Section Names

“references”, “notes”, “footnotes”, “sources”, “citations”, “works cited”, “bibliography”, “external references”, “reference notes”, “references cited”, “bibliographical references”, “cited references”, “see also”, “notes, references”, “sources, references, external links”, “sources, references, external links, quotations”, “notes & references”, “references & notes”, “external links & references”, “references & external links”, “references & footnotes”, “footnotes & references”, “citations & notes”, “notes & sources”, “sources & notes”, “notes & citations”, “footnotes & citations”, “citations & footnotes”, “reference & notes”, “footnotes & sources”, “note & references”, “notes & reference”, “sources & footnotes”, “notes & external links”, “references & further reading”, “sources & references”, “references & sources”, “references & links”, “links & references”, “references & bibliography”, “references & resources”, “bibliography & references”, “external articles & references”, “references & citations”, “citations & references”, “references & external link”, “external link & references”, “further reading & references”, “notes, sources & references”, “references and further reading”, “sources, references & external links”, “references/notes”, “notes/references”, “notes/further reading”, “references/links”, “external links/references”, “references/external links”, “references/sources”, “external links / references”, “references / sources”, “references / external links”

Bibliography

[1] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *J. of Management Information Systems*, vol. 12, no. 4, pp. 5–33, 1996.

[2] J. Juran, *Juran on Quality by Design.* Free Press, 1992.

[3] K. Ivanov, *Quality Control of Information : On the Concept of Accuracy of Information in Data-banks and in Management Information Systems.* The Royal Institute of Technology, Department of Information Processing Computer Science, Stockholm, 1972.

[4] B. Leuf and W. Cunningham, *Methods of Social Research.* Addison-Wesley, 2001.

[5] P. Dondio and S. Barrett, "Computational trust in web content quality: A comparative evalutation on the wikipedia project," *Informatica (Slovenia)*, vol. 31, no. 2, pp. 151–160, 2007.

[6] A. Halfaker, A. Kittur, R. Kraut, and J. Riedl, "A jury of your peers: quality, experience and ownership in wikipedia," in *Int. Sym. Wikis*, D. Riehle and A. Bruckman, Eds. ACM, 2009.

[7] A. Kittur and R. E. Kraut, "Harnessing the wisdom of crowds in wikipedia: quality through coordination," in *CSCW*, B. Begole and D. W. McDonald, Eds. ACM, 2008, pp. 37–46.

[8] A. Lih, "Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource," *ISOJ*, pp. 1–31, 2004.

[9] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser, "Information quality work organization in wikipedia," *JASIST*, vol. 59, no. 6, pp. 983–1001, 2008.

[10] D. M. Wilkinson and B. A. Huberman, "Cooperation and quality in Wikipedia," in *Int. Sym. Wikis*, A. Désilets and R. Biddle, Eds. ACM, 2007, pp. 157–164.

[11] J. Giles, "Internet encyclopaedias go head to head," *Nature*, vol. 438, pp. 900–901, 2005.

[12] D. H. Dalip, M. A. Gonçalves, M. Cristo, and P. Calado, "Automatic assessment of document quality in web collaborative digital libraries," *J. Data and Information Quality*, vol. 2, no. 3, p. 14, 2011.

[13] B. Stvilia, M. B. Twidale, L. Gasser, and L. C. Smith, "Information quality discussion in Wikipedia," ISRN UIUCLIS, Tech. Rep., 2005.

[14] K. Bailey, *Methods of Social Research*. New Press, 1994.

[15] J. E. Blumenstock, "Size matters: Word count as a measure of quality on Wikipedia," in *WWW*, J. Huai, R. Chen, H.-W. Hon, Y. Liu, W.-Y. Ma, A. Tomkins, and X. Zhang, Eds. ACM, 2008, pp. 1095–1096.

[16] ——, "Automatically assessing the quality of Wikipedia articles encyclopedia," University of California at Berkeley School of Information, Tech. Rep., 2008.

[17] N. Lipka and B. Stein, "Identifying featured articles in Wikipedia: Writing style matters," in *WWW*, M. Rappa, P. Jones, J. Freire, and S. Chakrabarti, Eds. ACM, 2010, pp. 1147–1148.

[18] T. M. Mitchell, *Machine Learning*. McGraw-Hill, 1997.

[19] M. Anderka, B. Stein, and N. Lipka, "Predicting quality flaws in user-generated content: the case of wikipedia," in *SIGIR*, W. R. Hersh, J. Callan, Y. Maarek, and M. Sanderson, Eds. ACM, 2012, pp. 981–990.

[20] M. Hu, E.-P. Lim, A. Sun, H. W. Lauw, and B.-Q. Vuong, "Measuring article quality in Wikipedia: Models and evaluation," in *CIKM*, M. J. Silva, A. H. F. Laender, R. A. Baeza-Yates, D. L. McGuinness, B. Olstad, Ø. H. Olsen, and A. O. Falcão, Eds. ACM, 2007, pp. 243–252.

[21] D. H. Dalip, M. A. Gonçalves, M. Cristo, and P. Calado, "Automatic quality assessment of content created collaboratively by web communities: A case study of Wikipedia," in *JCDL*, F. Heath, M. L. Rice-Lively, and R. Furuta, Eds. ACM, 2009, pp. 295–304.

[22] E. Smith and R. Senter, "Automated readability index," AMRL-TR-66-220, Aerospace Medical Research Laboratories, Wright-Patterson Airforce Base, OH., Tech. Rep., 1967.

[23] M. Coleman and T. L. Liau, "A computer readability formula designed for machine scoring," *Applied Psychology*, pp. 283–284, 1975.

[24] R. Flesch, "A new readability yardstick," *Applied Psychology*, pp. 221–233, 1948.

[25] J. Kincaid, R. Fishburne, R. Rogers, and B. Chissom, "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel," Naval Technical Training Command, Millington, TN. Research Branch, Tech. Rep., 1975.

[26] J. S. Caylor and T. G. Sticht, "Development of a simple readability index for job reading material," Human Resources Research Organization, Monterey, CA. Div. 3., Tech. Rep., 1973.

[27] R. Gunning, *The technique of clear writing.* McGraw-Hill, 1952.

[28] C.-H. Björnsson, *Läsbarhet.* Pedagogiskt Centrum, 1968.

[29] G. H. McLaughlin, "SMOG grading - a new readability formula," *Journal of Reading*, pp. 639–646, 1969.

[30] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks*, vol. 30, no. 1-7, pp. 107–117, 1998.

[31] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser, "Assessing information quality of a community-based encyclopedia," in *IQ*, F. Naumann, M. Gertz, and S. E. Madnick, Eds. MIT, 2005, pp. 442–454.