J Intell Inf Syst (2017) 48:601–632
DOI 10.1007/s10844-016-0428-1

# Diversity of editors and teams versus quality of cooperative work: experiments on wikipedia

**Marcin Sydow[1,2] · Katarzyna Baraniak[1,2] · Paweł Teisseyre[2]**

Received: 19 January 2016 / Revised: 21 August 2016 / Accepted: 23 August 2016 /
Published online: 7 October 2016


**Abstract** We study whether and how the diversity of editors and teams affects the quality of work in a virtual cooperative work environment on the Wikipedia example. We propose a measure of interests diversity of an editor and some measures of team diversity in terms of members' interests and experience. Statistical and machine learning methods are used to investigate the dependency between diversity and work quality. The presented experimental results confirm our hypothesis that interest diversity of a single editors and team diversity are positively related to the quality of their work. Interestingly, some of our experiments also indicate that diversity may be more important than such attributes as productivity of an editor or size or experience of the team. Our experimental results demonstrate that it is possible to predict work quality based on diversity which is an additional statistical signal that diversity is correlated with work quality.

**Keywords** Diversity of interest · Team diversity · Wikipedia · Article quality · Open collaboration · Machine learning

## 1 Introduction

Common access to the Internet made it possible that virtual open-collaboration environments became an important platform for massive collaborative work. A good example is

✉ Marcin Sydow
msyd@poljap.edu.pl

Katarzyna Baraniak
katarzyna.baraniak1@pjwstk.edu.pl

Paweł Teisseyre
pawel.teisseyre@ipipan.waw.pl

[1] Polish-Japanese Academy of Information Technology, Warsaw, Poland

[2] Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

Wikipedia, where editors work on preparing articles. However, the quality of such work significantly varies between particular articles, editors and teams of editors working together on articles. It is important to study which factors of an editor or team influence the quality of the outcome of such collaborative work. For example, it is interesting to study whether an editor that has *diverse* interests (i.e. is "versatile") tends to create better Wikipedia articles. It is even more interesting whether teams that are diverse in terms of interest or experience of their members tend to produce better articles.

In this article we study whether and how the interests diversity of editors and interest and experience diversity of editor teams affect the quality of work in a virtual cooperative work environment on the Wikipedia example. In future, such studies can help to develop and improve the tools supporting open-collaboration team-building process.

Diversity has proved to play an important role in multiple fields of information sciences and applications such as: text summarisation, web search (Agrawal et al. 2009), databases (Vee et al. 2008), recommender systems and semantic entity summarisation (Sydow et al. 2013). Recent research also indicates that diversity of population plays a positive role in evolutionary algorithms (Strzezek et al. 2015)

Our hypothesis studied in this article is that *diversity of editors and teams is a factor that positively affects the quality of work* in a virtual cooperative environments.

To verify this hypothesis experimentally we statistically analyse data from the Polish and German Wikipedia.

We introduce several quantitative measures of diversity of a member of an open-collaboration environment or a whole team thereof. One of the proposed measures is based on the information-theoretic concept of *entropy*, whereas other measures are based on statistical standard deviation.

In order to study how these measures influence the work quality we use statistical and machine learning techniques, which are very effective tools to investigate such dependencies. We demonstrate on Wikipedia data that interest diversity of an editor seems to be correlated with the quality of the articles they co-edit. We also extend the concept of interest diversity on whole teams of authors and study how it impacts the work quality compared to their productivity and experience. In the case of teams the reported experimental findings are similar: team's diversity is correlated with quality.

We also demonstrate that it is possible to use statistical machine learning tools to predict the quality of Wikipedia articles using some attributes that model the level of editors' diversity (and some other attributes) which can be interpreted as an additional statistical signal that diversity positively affects work quality in Wikipedia.

### 1.1 Motivation

Team diversity is one of the fundamental issues in social and organisational studies that has been broadly researched on (e.g. Parnas 1972; Sanchez and Mahoney 1996; Langlois and Garzarelli 2008). It has been broadly theorised and tested on virtual communities. One of the most burning questions concerns team coherence vs efficiency. There are two competing theories describing the efficient team organisation: modularity and integrity (Parnas 1972; Sanchez and Mahoney 1996). The first was introduced by David Parnas who suggested that co-dependence between "components" or "modules" (in our context this concept corresponds to an article on Wikipedia) should be eliminated by limiting the communication induced by the modules (Parnas 1972). In this approach participation in a module does not require knowledge about the whole system or other modules, e.g., Wikipedia users can co-author articles about social science without knowing anything about life sciences or

mathematics. It leads to higher specialization and less diversity in individual performance. A modular approach enables more flexibility and decentralized management (Sanchez and Mahoney 1996).

In the integral mode the team members have diverse knowledge and skills. We aim to study whether modular/specialized or integral collaboration pattern is more successful in creating high-quality Wikipedia articles.

### 1.2 Contributions

The contributions of our work include:

- the concept of editor's "versatility" (interest diversity) based on information entropy and various measures of team diversity based on editor's versatility and statistical standard deviation of selected attributes,
- exploratory analysis of two datasets based on dumps of Wikipedia (Polish and German), which indicate that versatility of editors and diversity of teams is positively correlated with quality of articles,
- exploratory analysis of relationship between editor's gender and versatility,
- more sophisticated statistical analysis of the studied datasets that includes a series of experiments with various machine learning prediction algorithms (logistic regression, decision trees) that verify whether and how accurately it is possible to predict the quality of articles based on some characteristics of their editors with special focus on diversity,
- analogous series of experiments concerning teams of editors, applying logistic regression and random forests,
- additional analyses utilising importance measures that further support the thesis that diversity is the most important factor in the presented prediction models,
- additional analysis in the form of various graphs concerning the performance of the prediction models (Lift and ROC curves) that further support the previous findings.

This article is a substantial extension of a conference paper (Baraniak et al. 2016) where the parts of the two first of the above contributions were preliminarily presented.

Our experimental results seem to positively confirm hypothesis that diversity of single editors and teams is positively related to the quality of their work and that diversity is usually more important than some seemingly more obvious attributes such as size or productivity of the team.

### 1.3 Related work

The general comparison of quality of classic and open-collaboration encyclopediae, in particular Britannica vs Wikipedia is discussed in Giles (2005) when it is observed that the quality of Wikipedia (in terms of number of errors) is not much lower than that of Britannica, which is a bit surprising result.

The problem of how the number of editors and the coordination method of their work influences the article quality is studied in Kittur and Kraut (2008). Two coordination methods are considered: the explicit one and the implicit one. In the latter one, the work is planned and coordinated by explicit communication between all the editors while in the second one the most of the work is organised and done by a small subset of the editor team. The presented results demonstrate that adding more editors can improve the article quality only if the applied work coordination method was appropriate. In particular the results indicate that the implicit coordination helps more in larger teams.

The interplay between the phenomenons of social influence and social preference based on similarity between the editors in the context of open collaboration in Wikipedia is studied in Crandall et al. (2008). The results indicate that both phenomenons play an important role in explaining the open collaboration patterns.

In Wilkinson and Huberman (2007) it is reported that the high-quality articles are those that are intensively edited and have high number of editors as compared to other articles of similar age. Our work shows that diversity is not less important in this context.

The important role of a diversity was noticed early not only in complex systems but also in other fields like Operation Research or Information Retrieval (Goffman 1964). One of the earliest successful applications of diversity-aware approach was reported in Carbonell and Goldstein (1998) in the context of text summarisation. Recently, diversity-awareness has gained increasing interest also in other information-related areas where the actual information need of a user is unknown and/or the user query is ambiguous so that a controlled level of diversity introduced to the results increases their quality. Examples range from databases (Vee et al. 2008) to Web search (Agrawal et al. 2009) or to the quite novel problem of graphical entity summarisation in semantic knowledge graphs (Sydow et al. 2013). A recent work (Strzezek et al. 2015) demonstrates that a controlled level of population diversity increases the performance of genetic algorithm for some hard optimisation problems.

The concept of diversity has also attracted interest also in the domain of open collaboration research, e.g. in Aggarwal (2014). From the open collaboration point of view, diversity can be considered from many perspectives, for example as a team diversity vs homogeneity or a single editor's versatility (called "integrity" in that work) vs specialisation (called "modularity" in that work).

The positive role of team diversity was studied in Chen et al. (2010), where productivity and diversity of teams can be defined in a different sense and it is suggested that other variables may have influence on the quality of article.

In our work we use different definitions of diversity and its measures, since we quantify it with the use of the concept of *entropy* and on standard deviation, as will be explained in Section 2. Most importantly, in contrast to our work, the mentioned work studies the influence of diversity on the amount of accomplished work and withdrawal behaviour rather than the work quality that is considered here.

In contrast to our work most of previous works focus on diversity of editor teams in terms of categories such as culture, ethnicity, age, etc. López and Butler (2013) studies how the content diversity influences on-line public spaces in the context of local communities. A recent example, with a special emphasis on ad-hoc "swift" teams where the members have very little previous interactions with each other is Aggarwal (2014). Vasilescu et al. (2015) studied gender diversity relationship with work outcome.

A recent article (Ren et al. 2015) studies how tenure diversity and interest variety affect group productivity and member withdrawal and how the two types of diversity evolve over time. The results of this work seem to indicate the importance of the interest and experience diversity in online collaboration but does not directly address the issue of how it impacts the *quality* of the resulting articles that is the topic of this article.

## 2 Measures of diversity

In this article, in order to objectively measure how diversity affects work quality, we introduce and apply some *measures of diversity*.

The first diversity measure that we propose for editors, *versatility*, is based on *information entropy* (Shannon 1948) that is commonly used in various domains as a natural measure of diversity. Here it is used to model interest diversity of a single actor of a cooperative network. In this measure we assume that there are available some *topical categories* in the collaborative work model. The versatility measure is described in Section 2.1.

We also use some other measures of diversity in our experiments concerning teams of editors, that are based on *standard deviation*. It is one of the statistical concepts that measures how much an attribute *varies* around its mean value and can also be considered as a natural choice for a diversity measure. We use standard deviation in our experiments concerning *teams* of actors of a collaborative network. We briefly remind the concept of standard deviation in Section 2.2

## 2.1 Versatility (measure of interest diversity)

In this section we explain the model of interest diversity that we apply in our approach. We use Wikipedia terminology to illustrate the concepts, however, our model can be adapted to other, similar open-collaboration cooperative work environments.

Let $X$ denote a group of Wikipedia editors. Editors participate in editing Wikipedia articles. Each article can be mapped to one or more categories from a pre-defined *set of categories* $C = \{c_1, \ldots, c_k\}$ that represent topics.

Each editor $x \in X$ in our model is characterised by his/her editing activity i.e., all editing actions done by $x$. We assume that the interests of an editor $x$ can be represented by the amount of work that $x$ committed to articles in particular categories.

Let $t(x)$ denote the total amount of textual content (in bytes) that $x$ contributed to all articles co-edited (up to the moment of doing the analysis) and let $t_i(x)$ denote the total amount of textual content that editor $x$ contributed to the articles belonging to a specific category $c_i$.[1]

Now, lets introduce the following denotation: $p_i(x) = t_i(x)/t(x)$ and interpret it as representing $x$'s *interest in category* $c_i$. Henceforth, we will use a shorter denotation $p_i$ for $p_i(x)$ whenever $x$ is understood from the context.

### 2.1.1 Editor's interest profile

Finally, we define the *interest profile* of the editor $x$, denoted as $ip(x)$, as the *interest distribution vector* over the set of all categories:

$$ip(x) = (p_1(x), \ldots, p_k(x)) \tag{1}$$

Notice that according to the definition the interest profile represents a valid distribution vector i.e., its coordinates sum up to 1.

### 2.1.2 Example

Assume that the set of categories $C$ consists of 8 categories: $\{c_i\}_{1 \leq i \leq 8}$ and that editor $x$ has contributed $t(x) = 10kB$ of text in total, out of which $t_2(x) = 8kB$ of text has been

[1] Since a single article can be assigned to multiple categories, we split the contribution equally for all the categories of the article.

contributed to articles in category $c_2$, $t_5(x) = 2kB$ in category $c_5$ and nothing to articles that were not assigned to $c_2$ nor $c_5$. Thus, $x's$ interest in $c_2$ is $p_2(x) = t_2(x)/t(x) = \frac{4}{5}$, in $c_5$ is $p_5(x) = t_5(x)/t(x) = \frac{1}{5}$ and is equal to 0 for all other categories. The interest profile of this user is:

$$ip(x) = (0, \frac{4}{5}, 0, 0, \frac{1}{5}, 0, 0, 0).$$

#### 2.1.3 Editor's versatility measure

There are many possible ways of measuring diversity. Since the interest profile $ip(x)$ is modelled as a distribution vector over categories, we define *diversity of interests* (or equivalently *versatility*) of $x$, $V(x)$, as the *entropy of interest profile* of $x$:

$$V(x) = H((p_1, p_2, \ldots, p_k)) = \sum_{1 \le i \le k} -p_k \log_2(p_k) \tag{2}$$

The value of entropy ranges from 0 which represents *extreme specialisation* (i.e. total devotion to a single category) to $log_2(k)$ which represents extreme diversity (i.e. active and equal interest in all possible categories).

Information entropy has several elegant and natural mathematical properties (Shannon 1948) and is a commonly used measure of diversity in various applications concerning information sciences.

#### 2.1.4 Example, continued

The versatility of user $x$ from Section 2.1.2 has the following value:

$$V(x) = -p_2 lg(p_2) - p_5 lg(p_5) = 0.8 \times 0.32 + 0.2 \times 2.32 = 0.256 + 0.464 = 0.72$$

Now assume that another user $x'$ has contributed equally to the four first categories, i.e. user's interest profile is: $ip(x') = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, 0, 0, 0, 0)$. The versatility value for this editor has the following value:

$$H(ip(x')) = -4 \times 0.25 \times (log_2(0.25)) = 2$$

Notice that the versatility measure of $x'$ is higher than that of $x$ and that this is according to the intuition since $x'$ has similar interest in four different categories and $x$ only in two (mostly in one). In other words, $x'$ is more versatile while $x$ is more specialised. Maximum versatility for $n$ categories would have value of $log_2(n)$, for an editor that is equally interested in all categories.

The datasets that are experimentally studied later in this article consider 8 and 12 categories, respectively, so that maximum versatility (entropy) for these cases would be $log_2(8) = 3$ and $log_2(12) \approx 3.584$, respectively.

### 2.2 Standard deviation

In this paper we also use some measures of diversity based on standard deviation. Standard deviation of numerical attribute $X$ taking $n$ values: $X_1, \ldots, X_n$ is defined as

$$\text{sd(X)} := \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \text{avg}(X))^2},$$

where $\text{avg}(X) = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an arithmetic mean of attribute $X$. Standard deviation sd(X) measures how much (on average) an attribute varies around its arithmetic mean. Thus it can be seen as a natural measure of variability or dispersion of a numerical attribute. In our experiments, we will use standard deviations of the number of editors' contributions in bytes and standard deviations of period lengths between the first and the last contributions (tenure, that may represent the experience of an editor).

## 3 Data

To verify our hypothesis in this article, i.e. to study the relationship between diversity and work quality in collaborative environments we apply experimental statistical analysis method to real data concerning collaborative work.

We decided to focus on one of the most popular environments of open collaborative work – Wikipedia, since it is quite large, rich in attributes, publicly available and, in addition, provides means of measuring *quality* of the work.

### 3.1 Data mining approach to the problem

We prepared datasets and preprocessed them to compute several attributes for editors and teams of editors. We also utilised information available in Wikipedia to attach a *quality label* to each article that is treated as the *decision attribute* in our analyses.

We first run statistical tools to preliminarily statistically analyse the relationship between diversity and other attributes and quality.

Next, for a deeper analysis, we additionally applied some more sophisticated statistical machine learning tools such as logistic regression, decision trees, random forests. The methods are described in more detail in Section 4.

In such models it is possible to objectively measure in various ways how strongly any attribute is correlated with the decision attribute (quality in our case). In particular, in some experiments we split our preprocessed data into *training* and *test* sets, built *prediction* models based on them and used the models *to predict* work quality based on the studied attributes.

The higher performance of such prediction, the stronger statistical relationship between the attributes (including diversity) and the decision attribute (quality). In addition we applied some other statistical tools to objectively measure how strongly diversity (and other attributes) affects the quality of work.

### 3.2 Datasets

The activity of editors and their teams on Wikipedia are recorded and stored in Wikipedia dumps that are publicly and easily available. Wikipedia shares the latest dumps under the following URL address: https://dumps.wikimedia.org/.

To run our experimental study, we prepared ourselves two separate datasets by processing dumps of the Polish and German Wikipedia from March and September of 2015, respectively. We will refer to these two datasets as `wiki-pl` and `wiki-de`, respectively.[2]

[2]Datasets used in this article for the experiments are available on e-mail request.

We run all the experiments presented in this article on two different language versions of Wikipedia for greater reliability of the results.

Since the results (presented later on in this article) on both datasets wiki-pl and wiki-de are generally compatible, we assume that the choice of these particular language versions of Wikipedia does not significantly affect our general findings presented in this article.

We collected data about editors of articles, articles and editions of articles made by authors.

By *edition* we mean any contribution of an editor to an article that results in the change of the article's content by editing it (for example: adding a content by inserting new paragraph or modyfying an existing paragraph, etc.).

The datasets used in our article are summarised in the Table 1.

### 3.3 Means of measuring the quality of wikipedia articles

In this article we use the information assigned explicitly to Wikipedia articles to determine their *quality*. More precisely, the quality of articles is modelled based on the information given by Wikipedia community members, who evaluate articles as *good* and/or *featured* based on the following criteria explicitly defined by the Wikipedia community itself:

- *GOOD* article (G): *"well-written, comprehensive, well-researched, neutral, stable, illustrated"*
- *FEATURED* article (F): (in addition to the above) *"length and style guidelines including a lead, appropriate structure and consistent citation"*

We utilise the above two kinds of labels given by the Wikipedia community regarding the articles as the basic means of determining their quality.

Based on this, in our considerations and experiments we distinguish five quality classes of editors denoted as $N$, $G \cup F$, $G \cap F$, $G$, $F$ as presented in Table 2.

**Note** As we define the class denoted as $G$ as the class of editors who edited at least one good article *and no featured article* and analogously the class $F$, the more obvious denotations for these classes would actually be $G \setminus F$ and $F \setminus G$, respectively. However we use $G$ and $F$ to simplify the notation. Notice that this simplification implies that classed denoted as $G$, $F$, $G \cap F$ are actually *mutually exclusive* and they *split* the $G \cup F$ class into three different sub-classes.

It is natural to observe that the introduced editor quality classes exhibit partial order "hierarchy" among the editors. In such interpretation the $G \cap F$ represents the highest-quality editors and $N$ the lowest, etc.

Table 3 presents the sizes of all considered quality classes in our datasets. Notice that in our datasets all articles can be marked either "good" or "featured" (not both at the same time), however an editor (author) can contribute to articles representing both classes.

**Table 1** Summary of Datasets wiki-pl and wiki-de, *"edition"* is any contribution of an editor to an article that results in the change of the article's content

| | wiki-pl dataset | wiki-de dataset |
|---|---|---|
| Editors | 126,406 | 555,355 |
| Articles | 947,080 | 1,422,940 |
| Editions | 16,084,290 | 61,266,990 |

**Table 2** Analysed quality groups of editors

| Editor quality class | Definition |
|---|---|
| N | (normal) edited *no good nor featured* article |
| G∪F | (good or featured) *at least one good or one featured* article |
| $G$ (denotes: $G \setminus F$) | (good) edited *at least one good article and no featured* article |
| F (denotes: $F \setminus G$) | (featured) edited *at least one featured article and no good* article |
| G∩F | (good and featured) edited *at least one good and one featured* article |

### 3.4 Topical categories of articles

Our definition of versatility (topical diversity) of an editor presented in Section 2.1 assumes the existence of topical categories.

In our model we utilised 12 main content categories for Polish Wikipedia and 8 main content categories for German Wikipedia accessible from the front page. We preprocessed our datasets wiki-pl and wiki-de to identify main categories assigned to considered articles. Table 4 presents the main categories for both datasets (they differ for language versions of Wikipedia).

Wikipedia articles are usually not directly tagged with any of these high-level categories. Only the most specific categories are assigned to the articles by Wikipedia community. Those are subcategories of more general categories, creating a structure of a directed graph. The nodes in this graph are categories and there is a directed arc from one vertex to another in such a graph if and only if the corresponding category is a subcategory of another. Starting from any node in the graph representing a lowest-level category directly assigned to a particular article, we employed a standard BFS (breadth-first search Cormen et al. 2001) graph search algorithm to assign top-level categories to this article. More precisely, the article was assigned all top-level categories reachible from the lowest-level category of this article by the BFS algorithm.

If the article was mapped to more than one category, the contribution size was split equally among them. Articles that couldn't be classified were excluded from the dataset, as well as users whose production consisted of such articles exclusively. Also, only editions of the pages in the primary namespace were taken into account (that is "proper" articles and not, for example, discussion pages), because only these pages are evaluated with regard to their quality.

**Table 3** Sizes of articles and editors among quality classes for wiki-pl and wiki-de datasets

| Number of | wiki-pl | wiki-de |
|---|---|---|
| Normal articles | 944,585 | 1,417,318 |
| Good articles | 1,889 | 3,424 |
| Featured articles | 606 | 2,198 |
| Editors of normal articles (N) | 124,673 | 479,908 |
| Editors of good articles and no featured articles ($G$) | 4,534 | 34,063 |
| Editors of featured articles and no good articles (F) | 2,272 | 17,797 |
| Editors of good or featured articles ($G \cup F$) | 9,939 | 75,447 |
| Editors of good and featured articles ($G \cap F$) | 3,133 | 23,587 |

**Table 4** Wikipedia main content categories

| Dataset | Main content categories |
|---|---|
| wiki-pl dataset | Humanities and Social Sciences |
| | Natural and Physical Sciences |
| | Art & Culture |
| | Philosophy |
| | Geography |
| | History |
| | Economy |
| | Biographies |
| | Religion |
| | Society |
| | Technology |
| | Poland |
| wiki-de dataset | Art & Culture |
| | Geography |
| | History |
| | Knowledge |
| | Religion |
| | Society |
| | Sport |
| | Technology |

### 3.5 Attributes of an editor

One of the main objects of study in our work is an editor of a Wikipedia article, i.e. a person who contributed to the work on the article. Size of editors contribution to article was counted as a sum of his editions to this article over all revisions in dump. We considered edition as size of one change by editor in one article. Wikipedia does not provide exact size of an edit and contribution so we had to count it. Every revision in a dump has an information about current size of text in bytes. Size of one edition made by editor was counted as difference of size between the last and current revision. Contribution is the amount of bytes changed by one editor in article. Datasets components presented in Table 5 was used to compute data for this part of the experiments. Anyone who made any change to article was treated as an editor even if it was just minor change like adding a comma, because criteria of good and featured articles include style and well-written.

Additionally, we gathered data about the editors' gender that is available in our datasets. Not all editors share this information on their Wikipedia profiles, but enough information was available to perform some basic analysis.

The sampling frame (observation interval) was restricted to contributors who made at least one edition during the Wikipedia project lifetime.

### 3.6 Additional data preparation for experiments with teams

In Section 6 we will present a series of experiments that will concern whole teams of editors.

**Table 5** Datasets for editors

| Components of the dataset | Description |
|---|---|
| Basic articles categories | Article id, article basic categories |
| Category graph | Category, more general categories of category |
| Main categories | Article id, main categories of article |
| Authors contributions | Contributor id, article id, size of contribution |
| Author versatility | Contributor id, contribution of author to main categories, versatility, the total size of edition made by author to all articles, flag if author contributes to good articles, flag if author contributes to featured articles |
| Authors gender | Contributor id, flag if author is woman, man or no information |

In our model we define *team*, associated to an article, as a group of all editors who contribute to this article. Our definition of team involves every editor who made any change within a particular article, such as text addition, deletion or some minor corrections. One editor may contribute to many articles but one team, according to our definition, creates only one article.

To perform team-oriented experiments, some further data processing was needed. We used wiki-pl and wiki-de datasets again. For each dataset, we precomputed three components of the data shown in Table 6, and integrated them into one dataset.

## 4 Statistical machine learning tools

In this section we describe machine learning models used in our experiments.

**Table 6** Datasets for teams

| Components of the dataset | Description |
|---|---|
| Editors edition | Contributor id |
| | Article id |
| | Size of edition (bytes) made by an editor to an article |
| | The total size of edition made by editor to all articles |
| Tenure of contributions | Contributor id |
| | Article id |
| | The number of days spent on article |
| | The number of days on Wikipedia |
| Diversity of interest | Article id |
| | Mean contribution of team members to main categoriesversatility of team |
| | The quality of article |

### 4.1 Logistic regression

Logistic regression (Hosmer et al. 2013) belongs to the most popular and successful classification models. Let $C$ be the value of a binary class variable and $x_1, \ldots, x_p$ be a set of numerical attributes. In logistic regression it is assumed that the posterior probability (the conditional probability of the class given attributes) is of the form

$$P(C = 1|x_1, \ldots, x_p) = \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots, +\beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots, +\beta_p x_p)},$$

where $\beta_0, \beta_1, \ldots, \beta_p$ are parameters. The parameters are usually estimated from data using a maximum likelihood method. The significance of the attribute in the model can be assessed using the Wald statistic (often denoted as $z$ statistic). The $z$ statistic for the $j$-th attribute is defined as the standardized estimator of the coefficient corresponding to the $j$-th attribute. The statistic is used to test the hypothesis $\beta_j = 0$. For a large sample size, $z$ statistic follows standard Gaussian distribution and thus the p-value of the statistic can be calculated. The smaller the p-value the more significant is the variable. Tables 10, 11, 16 and 17 contain values of $z$-statistics and the corresponding p-values from logistic regression.

### 4.2 Decision tree

Decision tree (Breiman et al. 1984) is an example of a non-linear classification model. In each node of the tree the data is split into two subsets according to the outcome of the test. The splits are performed in order to decrease the homogenity of the class distribution. The most popular measures of the class homogenity are: entropy and Gini index. The paths from root to leaves represent classification rules. The final decision is made based on the majority class in the given leaf. The optimal size of the tree can be determined using e.g. cost-complexity criterion. Figures 3 and 4 show the trees built based on our data.

### 4.3 Random forest

Random Forest (Liaw and Wiener 2002) consists of many single decision trees. Each of the tree is built based on boostrap sample (sample drawn with replacement from the original data). In addition, the Random Forest use a modified tree learning algorithm that selects, at each candidate split in the learning process, a random subset of the features. Random Forests corrects for a single decision trees' habit of overfitting to their training data. The final classification rule is based on the majority voting of the trees. Random Forest can be used to assess the importances of the attributes. The two basic measures are described in Section 6.6.

## 5 Experimental results for editors

In this section we report a series of experiments whose object of study is a single *editor*. More precisely, we experimentally study whether and how strongly versatility of an editor is correlated with the quality of the articles they co-edit.

In these experiments we measure the level of interest diversity of an editor with the versatility measure defined in Section 2.1.

The order of the experiments is as follows. In Section 5.1 we present a preliminary exploratory data analysis. We complete the exploratory analysis in Section 5.2 where we compare versatility of women and men to see whether the gender has any relationship with diversity of interest and quality.

Next, we present a deeper analysis of the problem by using some prediction models. We describe the experimental setup including split into training and testing sets in Section 5.3. We apply the logistic regression model to explain quality in Section 5.4. Next, Section 5.5 introduces prediction performance metrics such as precision, recall, F-measure that are used in the remaining experiments with prediction models such as logistic regression and trees. The prediction results are presented in Section 5.6. The analysis is completed with additional graphs presenting Lift and Roc curves in Section 5.7 to deeper understand the prediction experiments.

A short summary of the experimental results concerning editors is given in Section 5.8.

### 5.1 Preliminary exploratory analysis of the data

Initially, we make some basic analysis of versatility level across all quality classes that were defined in Section 3.3. The results in a form of *box-plots* are presented in Figs. 1 and 2, for wiki-pl and wiki-de datasets, respectively. The results are encouraging for further analyses, since one can see on these figures that higher quality editors (classes $F$, $G$, $F \cap G$, $F \cup G$) tend to be more versatile (i.e. of higher entropy of interests): in terms of median, lower and higher quartiles (horizontal bars of the boxes). The situation is very similar for both analysed language versions of data.

The precise values of median versatility across the quality classes are presented in Table 7, first column. These results preliminarily indicate that diversity of interest seems to be correlated with quality.

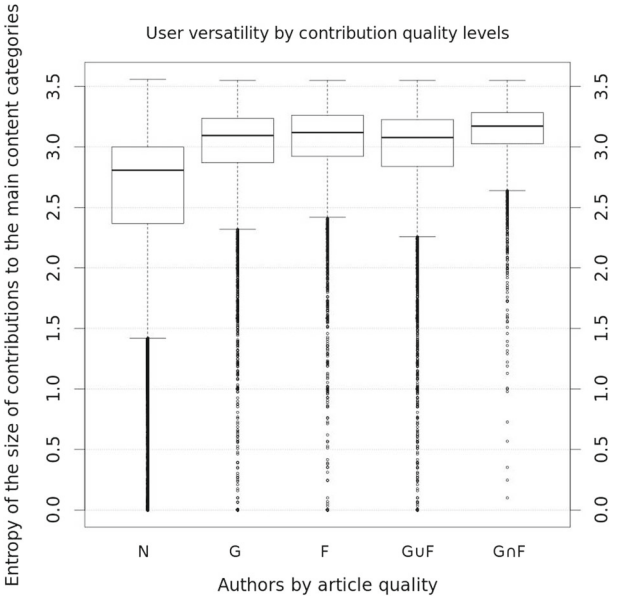


**Fig. 1** Versatility vs Quality for wiki-pl dataset

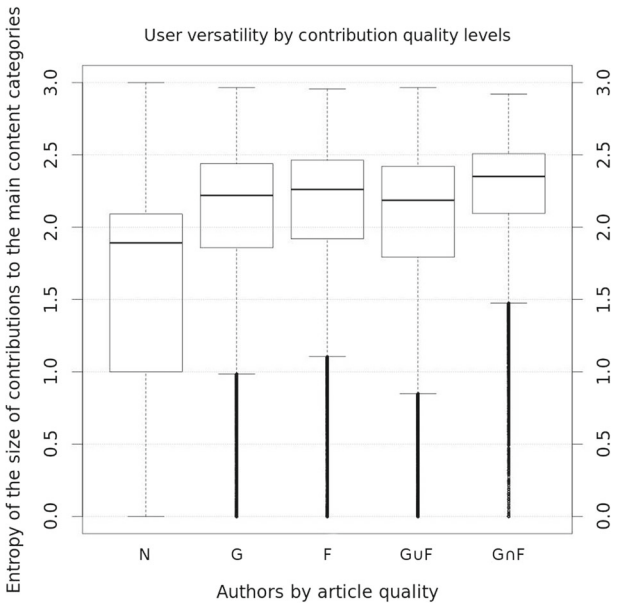


**Fig. 2** Versatility vs Quality for wiki-de dataset (denotations as on Fig. 1)

We also computed several other attributes of editors and preliminarily examined them against quality in order to compare them against versatility. One of the attributes that naturally comes in mind when analysing quality is *productivity* of an editor (total amount of work committed).

Indeed, our analysis confirmed that productivity is another editor's attribute that seems to be related to work quality. In the second column of Table 7 one can see that median productivity even stronger discriminates the quality classes than versatility.

As we will see in next experiments this seemingly superiority is misleading, since versatility better explains quality than productivity when more sophisticated statistical tools are applied.

Nonetheless, we selected productivity as the main "competitor" for versatility in the next experiments.

**Table 7** Median of versatility and productivity of editors vs. quality for wiki-pl and wiki-de dataset

| | wiki-pl | | wiki-de | |
|---|---|---|---|---|
| Quality | Versatility | Productivity | Versatility | Productivity |
| G∩F | 3.1720 | 159300 | 2.351 | 46080 |
| G∪F | 3.011 | 2992 | 2.064 | 1502 |
| F: | 3.000 | 2322 | 2.053 | 1283 |
| *G*: | 3.016 | 3347 | 2.070 | 1629 |
| N: | 2.807 | 237 | 1.891 | 264 |

**Table 8** Editors versatility vs gender (no observable relationship)

| Quality | Number of women | Number of men | Versatility of women | Versatility of men |
|---|---|---|---|---|
| wiki-pl | | | | |
| G∩F | 1.73e+02 | 3.98e+02 | 3.25e+00 | 3.25e+00 |
| G∪F | 2.46e+02 | 5.69e+02 | 3.18e+00 | 3.20e+00 |
| F: | 2.00e+01 | 4.70e+01 | 3.01e+00 | 3.02e+00 |
| *G*: | 5.30e+01 | 1.24e+02 | 3.09e+00 | 3.06e+00 |
| N: | 1.81e+02 | 4.14e+02 | 2.87e+00 | 2.91e+00 |
| wiki-de | | | | |
| G∩F | 5.53e+002 | 1.03e+003 | 2.51e+000 | 2.41e+000 |
| G∪F | 6.43e+002 | 1.32e+003 | 2.46e+000 | 2.44e+000 |
| F: | 3.40e+001 | 8.00e+001 | 2.17e+000 | 2.14e+000 |
| *G*: | 5.60e+001 | 2.11e+002 | 2.07e+000 | 2.18e+000 |
| N: | 1.95e+002 | 5.29e+002 | 1.84e+000 | 2.00e+000 |

Note: the classes $F$, $G$ and $G \cap F$ are mutually non-intersecting according to their definitions in Table 2

### 5.2 Exploratory analysis concerning the gender of editors

Some works study how gender relates to work outcome (e.g. Vasileseu et al. 2015). We split the editor data by their gender (Table 8, columns 1 and 2) for those editors who declared it and did an exploratory analysis concerning whether the relationship between versatility and quality differs between the both genders.

Comparison of editor versatility across all quality classes for both genders is presented in Table 8, columns 3 and 4, and indicates that there is no observable relationship between the gender and versatility across all classes. Versatility of women and men is more or less similar for each quality group and both examined datasets. We excluded gender factor from further experiments in this article.

### 5.3 Quality-prediction experimental setup

The remaining experiments aim at studying on how accurately it is possible to predict the quality group of the editor based on his/her versatility and productivity. Such approach of applying prediction models makes it possible to make a deeper analysis of the relationship between the examined attributes and quality. In general, higher prediction performance in such models may be interpreted as a statistical signal of dependence. In addition, various statistics concerning the prediction models such as: p-values, z-values, estimated coefficients, give more precise information about the relationship between the attributes that was not available in simple exploratory analysis.

For clarity, in this series of experiments, we set the prediction problem as the binary classification problem, where class variable $C = 1$ corresponds to G∪F ("high quality") editors, whereas class $C = 0$ corresponds to the remaining ones. The classes are heavily unbalanced (there are many fewer cases in class $C = 1$, see Tables 3 and 9), which makes the problem challenging from the statistical point of view.

We use two classification models, which are among the most popular ones in the machine learning community: logistic regression (Hosmer et al. 2013) and decision trees (Breiman et al. 1984). These two classifiers represent different groups of methods: the former one is

**Table 9** Class distributions for wiki-pl and wiki-de datasets

| wiki-pl | | wiki-de | |
|---|---|---|---|
| $C = 1$ | $C = 0$ | $C = 1$ | $C = 0$ |
| 9,939 | 134,612 | 75,447 | 555,355 |
| 6.87% | 93.12% | 11.96% | 88.03% |

an example of linear classifier, as the hyperplane separating the classes is a linear function of attributes. The latter model is a non-linear classifier. We use implementations available in the `R` (R Core Team 2013): the `glm{stats}` function for logistic regression and `rpart{rpart}` function for the tree (Therneau et al. 2015) (CART trees). Since the training data is unbalanced, we assign larger weights to articles from rare class when fitting a model.

To assess the predictive power of the considered methods, we randomly split our data into training (50 % observations) and testing (50 % observations) sets. The training data is used to build models (i.e. to fit logistic regression and build a decision tree), whereas testing data is used to check the prediction accuracy.

### 5.4 Explaining quality with logistic regression

The basic statistics presented in Table 7 indicate that productivity may be a factor as important as diversity in the context of work quality. To further examine this effect, we decided to use logistic regression to check how versatility, productivity and interactions between these two factors influence the quality group of editors which describes quality of his work and articles he edits. Tables 10 and 11 show the statistics from the fitted models: estimated coefficients (1st column), their standard errors (2nd column), Wald statistics, also called as "z-value" (3rd column) and the corresponding p-values (4th column). Small p-values indicate that the considered variables are statistically significant (i.e. are not "noisy"). High z-values indicate stronger relationship. Finally, the value of estimated coefficients objectively indicate how much one attribute affects another in the model. The sign of a coefficient represents the fact whether the influence is positive or negative. Row, corresponding to the most significant variable (we exclude the intercept coefficient as it is irrelevant in such consideration), is printed in bold. It is demonstrated that for both datasets *versatility is the most significant variable, representing the strongest relation and it positively affects the quality* in the model.

**Table 10** Logistic regression model predicting the quality group of editors on wiki-pl dataset. Interaction (product) of the variables is also included into the model

| | Estimate | Std. Error | z-value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | −5.35e+000 | 1.11e−001 | −48.115 | <2e-16*** |
| **versatility** | **9.32e-001** | **3.82e-002** | **24.384** | **<2e-16*** ** |
| productivity | −5.96e−006 | 2.74e−006 | −2.174 | 0.0297* |
| versatility×productivity (interaction) | 6.4e-006 | 9.18e-007 | 6.971 | 3.15e-012*** |

Signif. codes: p<0 '***', p<0.001 '**', p<0.01 '*', p<0.05 '.', p<0.1 ' '

**Table 11** Logistic regression model predicting the quality group of editors on wiki-de dataset. Interaction (product) of the variables is also included into the model

| | Estimate | Std. Error | z-value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | −3.539e+00 | 2.183e-02 | −162.110 | <2e-16*** |
| **versatility** | **7.879e-01** | **1.098e-02** | **71.767** | **<2e-16*** ** |
| productivity | 3.214e-06 | 5.829e-07 | 5.514 | 3.52e-08 *** |
| versatility×productivity (interaction) | 1.213e-05 | 3.317e-07 | 36.581 | <2e-16 *** |

Signif. codes: p<0 '***', p<0.001 '**', p<0.01 '*', p<0.05 '.', p<0.1 ' '

## 5.5 Prediction performance measures

In this section we remind some basic machine learning concepts that we use to further analyse our results in prediction experiments presented in the next Sections.

Let $C_i$ be the value of the class variable for the $i$-th observation in the test data and $\hat{C}_i$ be the predicted class for the $i$-th observation in the test data. Let's define the following quantities:

$$TP = |\{i : \hat{C}_i = 1 \text{ and } C_i = 1\}|,$$
$$FP = |\{i : \hat{C}_i = 1 \text{ and } C_i = 0\}|,$$
$$TN = |\{i : \hat{C}_i = 0 \text{ and } C_i = 0\}|,$$
$$FN = |\{i : \hat{C}_i = 0 \text{ and } C_i = 1\}|.$$

The letters 'T','F', 'P', 'N' denote "true", "false", "positive" and "negative", respectively. So, for example $TP$ ("true positive") is the number of cases correctly ("truly") assigned to class $C = 1$ ("positive"), etc. We use the following basic evaluation measures on the testing data:

$$\text{Precision} = \frac{TP}{TP + FP},$$
$$\text{True Positive Rate (Recall)} = \frac{TP}{TP + FN},$$
$$\text{F-measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision+Recall}}.$$

Precision measures how many articles are correctly predicted as $C = 1$ among those predicted as $C = 1$. Recall indicates how many articles are correctly predicted as $C = 1$ among all articles with label $C = 1$. In addition we calculate the F-measure, which is a harmonic mean of Recall and Precision. The higher the measures, the better the performance of the considered model. For highly unbalanced classes (such as in our data) it is hard to

**Table 12** Evaluation measures on testing data for editors on wiki-pl and wiki-de datasets

| Measure | Logistic regression wiki-pl dataset | Logistic regression wiki-de dataset | Tree model wiki-pl dataset | Tree model wiki-de dataset |
|---|---|---|---|---|
| Precision | 87.73% | 86.85% | 74.50% | 75.36% |
| Recall | 17.72% | 17.91% | 29.56% | 26.04% |
| Accuracy | 93.40% | 88.53% | 93.73% | 88.84% |
| F-measure | 29.48% | 29.70% | 42.33% | 38.70% |

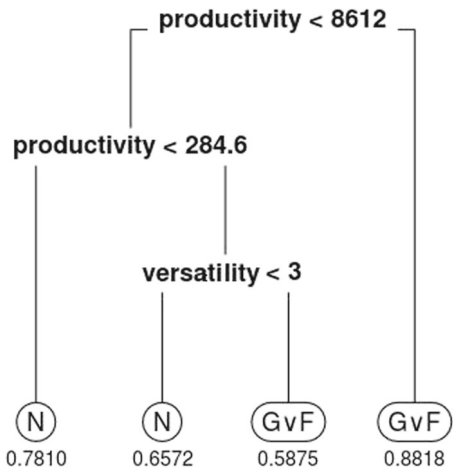


**Fig. 3** Tree model for wiki-pl dataset

build a model that optimises both Precision and Recall. Notice that for highly unbalanced classes the simple accuracy rate is misleading since it is easy to achieve its high value by always predicting the major class. Thus why F-measure as the aggregation of both is usually applied in such situations. The above indices are commonly used in machine learning and information retrieval.

### 5.6 Prediction results for logistic regression and tree model

In this section we present the performance measurements for editor quality prediction experiments (Table 12). Precision is higher for the prediction based on the logistic regression model (about $87-88$ %) than for the tree model ($74-75$ %), whereas recall is higher for the tree model ($26-29$ % for the tree model and 17 % for the logistic regression model). The tree model outperforms the logistic regression model with respect to F-measure, for both datasets.

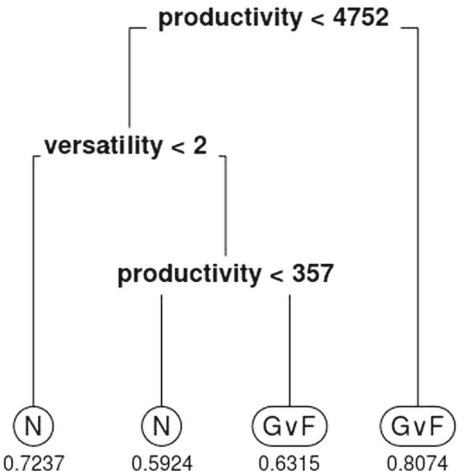


**Fig. 4** Tree model for wiki-de dataset

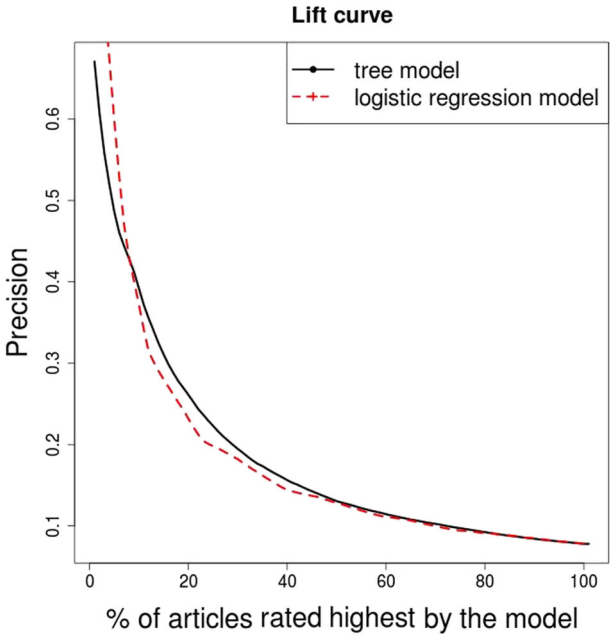


**Fig. 5** Lift curve for wiki-pl dataset models

Importantly, the results for wiki-pl and wiki-de datasets are quite similar which again supports the evidence that the choice of particular language version of Wikipedia does not affect the analysis.

Figures 3 and 4 present classification trees obtained in these experiments for wiki-pl and wiki-de datasets, respectively. They support the earlier observation that versatility is positively related to the high quality work or featured articles (class $C = 1$) as well as productivity.

In short, the presented prediction performance can be viewed as quite high if one takes into account high disproportion between class cardinalities. This can be interpreted as a signal of positive dependence between versatility and quality.

### 5.7 Lift and ROC curves

To complete the prediction-based analyses we present additional information about the prediction models that we obtained in our experiments.

This information is in the graphical form of *Lift curves* and *ROC curves* of the prediction models (Brown and Davis 2006). A lift curve graphically shows the precision (on the $y$-axis) with respect to the percentage of articles highest rated by the given model (on the $x$-axis). The precision in lift curve is calculated for the rule which assigns class $C = 1$ for articles highest rated by the given model (i.e. those with highest posterior probabilities).

ROC curve is another classical visualization tool, which shows the True Positive Rate (Recall) with respect to the False Positive Rate defined as follows:

$$\text{False Positive Rate} = \frac{FP}{FP + TN} \tag{3}$$

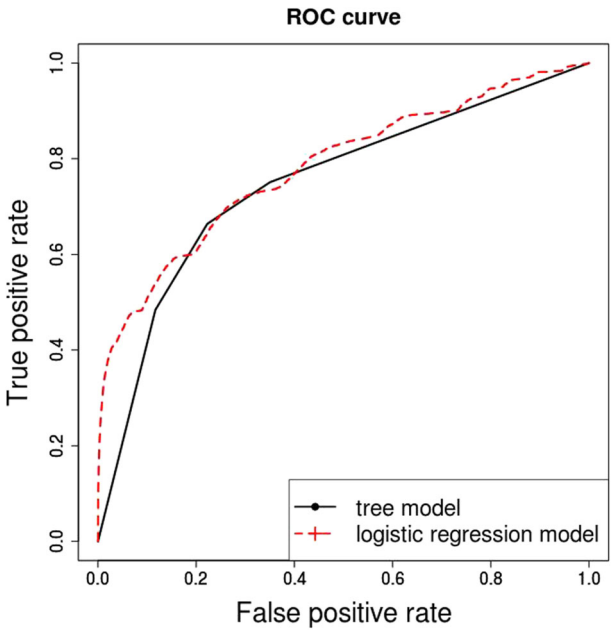


**Fig. 6** ROC curve for wiki-pl dataset models

(fraction of articles incorrectly predicted as $C = 1$ among those with label $C = 0$).

In general, the higher the area below the ROC curve, the better is the prediction model, with (ideal) maximum being 100 % of the area of the square.

Observe, that the baseline (random) assignment of articles to classes would result in a horizontal line on Lift chart and a diagonal line on ROC chart. The random assignment can be seen as a baseline (Figs. 5, 6, 7 and 8).

Figures 5 and 7 show Lift curves for the users of the wiki-pl and wiki-de dataset. Figures 6 and 8 show the corresponding ROC curves. To further explain the Lift curve graphs, observe that Fig. 5 indicates that when we assign class $C = 1$ to, for example, 10 % of observations highest rated by our models, we achieve precision about 40 %. Similarly, when we assign class $C = 1$ to say 40 % of observations that are highest rated by our models, we achieve precision about 15 %, etc.

Interestingly, both classification models give similar results. The results are promising as for both datasets we achieve the accuracy value which is definitely above the baseline (random assignment).

### 5.8 Summary of experimental results for editors

This Section summarizes the experiments concerning single editors. We used two statistical models to verify how much diversity (versatility) influences the quality group of Wikipedia editors. In addition, we also tested whether the productivity is correlated with the quality of an editors' group. It turns out that in both models, versatility is an important feature. In particular, versatility is the most significant variable according to the logistic model and it is also useful in the decision tree model. Analysis of an output from the logistic model indicates that versatility is positively correlated with the quality. Moreover, we tested the

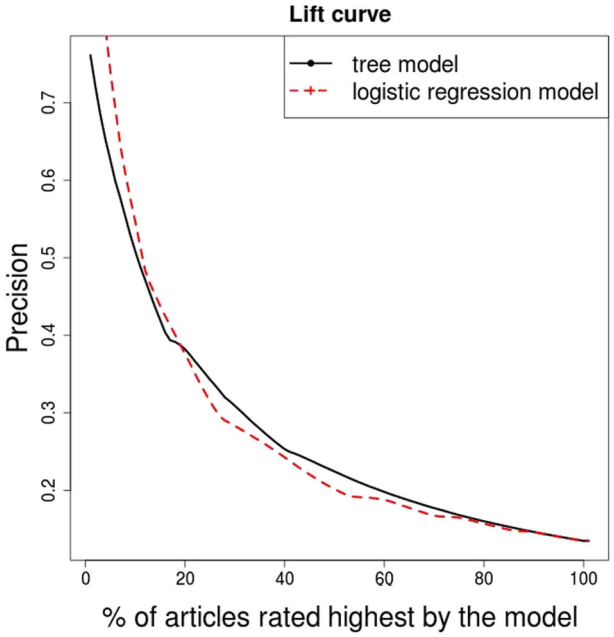


**Fig. 7** Lift curve for wiki-de dataset models

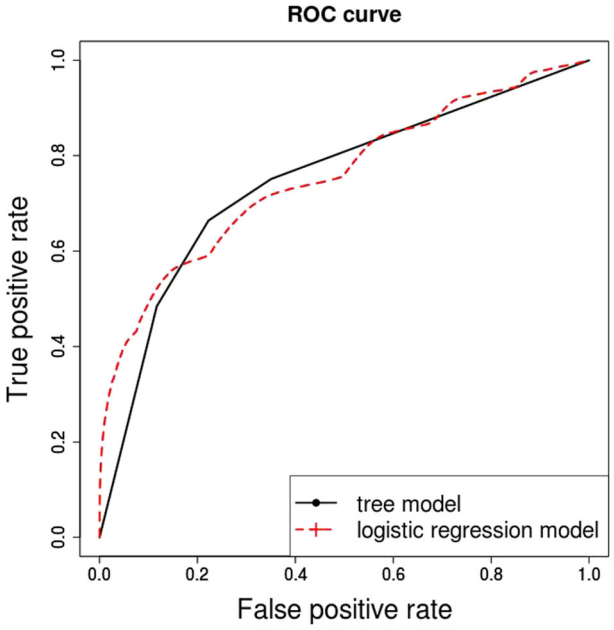


**Fig. 8** ROC curve for wiki-de dataset models

**Table 13** Attributes of teams

| Name | Description |
|---|---|
| Team size | $n = \|T\|$, where $T$ is a team i.e. a set of editors that work on a given article. |
| **Team versatility** | The *versatility* of a team $T$ is defined as the entropy of the *team interest profile* $tip(T)$ defined as follows. First, for each editor $x \in T$, we compute its individual interest profile $ip(x) = (p_1(x), \ldots, p_i(x), \ldots, p_k(x))$ as was defined in Section 2.1.1. Then, based on idividual interest profiles, for each topical category $i \in \{c_1, \ldots, c_k\}$ we compute the average (over team members) team interest in this category as $tp_i(T) = 1/n \sum_{x \in T} p_i(x)$ to form the team interest profile $tip(T) = (tp_1(T), \ldots, tp_i(T), \ldots, tp_k(T))$. Versatility is defined as entropy of this vector. |
| Mean productivity in the article | $MP(a) = \frac{1}{n} \sum_{i=1}^{n} P_i(a)$ is the mean amount of editors' contributions to the article $a$, where $P_i(a) = \sum_{e \in E_i(a)} \|newSize(e) - oldSize(e)\|$ is the total contribution of the $i-th$ editor to the article $a$, $E_i(a)$ is the set of the editions made by the editor $i$ in the article $a$ and $newSize(e)$, $oldSize(e)$ are the sizes of the article before and after the edition $e$, respectively |
| Mean total productivity | $MTP = \frac{1}{n} \sum_{i=1}^{n} TP_i$ is the mean amount of editors' contributions to all articles on the Wikipedia. Contribution is the sum of sizes (in bytes) of all editions made by team members to all articles in Wikipedia, where $TP_i = \sum_{e \in E_i} \|newSize(e) - oldSize(e)\|$ is the total contribution of the $i-th$ editor to all the Wikipedia articles and $E_i$ is the set of all the Wikipedia editions made by this editor |
| Mean tenure in article | $MT(a) = \frac{1}{n} \sum_{i=1}^{n} T_i(a)$ is the mean number of days spent on the article $a$ by the team members, where $T_i(a) = \|Df_i(a) - Dl_i(a)\|$ is the number of days between the date of the first $Df_i(a)$ and the last $Dl_i(a)$ date of any contribution of the $i-th$ editor to the article $a$ |
| Mean tenure in Wikipedia | $MTW = \frac{1}{n} \sum_{i=1}^{n} TW_i$ is the mean number of days spent on the Wikipedia, where $TW_i = \|DWf_i - DWl_i\|$, is the number of days between the first $DWf_i$ and the last $DWl_i$ date of any contribution of the $i-th$ editor to any Wikipedia article |
| **sd of productivity in article** | $SP(a) := \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (P_i(a) - MP(a))^2}$ the standard deviation of the $P_i(a)$ variable defined above |
| **sd of total productivity** | $STP := \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (TP_i - MTP)^2}$ the standard deviation of the $TP_i$ variable defined above |
| **sd of tenure in article** | $ST(a) := \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (T_i(a) - MT(a))^2}$ the standard deviation of the $T_i(a)$ variable defined above |
| **sd of tenure in wikipedia** | $STW := \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (TW_i - MTW)^2}$, the standard deviation of the $TW_i$ variable defined above |
| Length | $L(a)$ is the size of the article $a$ after the last recorded edition |
| Age | $AG(a) = Dc(a) - Dd$ the number of days between the date the article a was created and the date when dump was created |

A *team* (corresponding to an article) is a group of editors who contributed to any change in the article. The attributes related to the concept of *diversity* (as versatility, standard deviation, etc.) are printed in bold

prediction performance of these two models. The values of the applied evaluation measures (Precision, Recall, F-measure) are very promising. They are much larger than for the baseline (random assignment of articles to classes describing quality). This is also confirmed

**Table 14** Median of team features vs. quality articles of wiki-pl dataset

| Quality | Versatility | **Mean productivity in articles** | Mean total productivity | **sd productivity in articles** | sd total product. | **Length** |
|---|---|---|---|---|---|---|
| G∪F | 3.26e+000 | 1.80e+003 | 4.52e+006 | 6.84e+003 | 5.35e+006 | 3.19e+004 |
| F | 3.26e+000 | 2.93e+003 | 4.31e+006 | 9.62e+003 | 5.42e+006 | 5.38e+004 |
| *G* | 3.26e+000 | 1.73e+003 | 4.58e+006 | 6.10e+003 | 5.33e+006 | 2.70e+004 |
| N | 3.53e+000 | 4.99e+002 | 5.88e+006 | 7.96e+002 | 5.96e+006 | 2.41e+003 |
| Quality | **Team size** | **Mean tenure in article** | Mean tenure in Wikipedia | **sd tenure in article** | sd tenure in Wikipedia | Age |
| G∪F | 2.00e+001 | 1.25e+002 | 1.81e+003 | 3.56e+002 | 8.46e+002 | 2.59e+003 |
| F | 3.30e+001 | 1.44e+002 | 1.85e+003 | 4.11e+002 | 9.02e+002 | 3.13e+003 |
| *G* | 1.70e+001 | 1.20e+002 | 1.80e+003 | 3.37e+002 | 8.20e+002 | 2.43e+003 |
| N | 4.00e+000 | 7.71e+000 | 1.81e+003 | 4.39e+001 | 8.15e+002 | 2.31e+003 |

The attributes with observable discrimination potential between the *N* class against the other quality classes are printed in bold

by ROC and Lift curves. Both statistical models give comparable results and the conclusions are similar for both languages. The most important remark is that there is a strong dependence of versatility and work quality for editors.

## 6 Experimental results for teams

The experimental results, presented in Section 5, indicate that versatility of editors is positively dependent on the quality of their work. In this Section we extend the study to whole teams of editors and introduce many more attributes, including some new diversity-related ones.

**Table 15** Median of team features vs. quality articles of wiki-de dataset

| Quality | Versatility | **Mean product. in art.** | Mean total product. | **sd product. in art.** | **sd total product.** | **Length** |
|---|---|---|---|---|---|---|
| G∪F | 2.65e+000 | 1.16e+003 | 5.94e+006 | 6.05e+003 | 1.31e+007 | 4.28e+004 |
| F | 2.65e+000 | 1.44e+003 | 6.12e+006 | 8.09e+003 | 1.37e+007 | 5.58e+004 |
| *G* | 2.65e+000 | 9.98e+002 | 5.82e+006 | 4.98e+003 | 1.27e+007 | 3.58e+004 |
| N | 2.62e+000 | 4.07e+002 | 6.16e+006 | 9.10e+002 | 9.20e+006 | 3.64e+003 |
| Quality | **Team size** | **Mean tenure in article** | Mean tenure in Wikipedia | **sd tenure in article** | **sd tenure in Wikipedia** | Age |
| G∪F | 7.45e+001 | 1.02e+002 | 2.09e+003 | 3.33e+002 | 1.05e+003 | 3.74e+003 |
| F | 8.60e+001 | 1.01e+002 | 2.11e+003 | 3.30e+002 | 1.05e+003 | 3.83e+003 |
| *G* | 6.60e+001 | 1.03e+002 | 2.08e+003 | 3.36e+002 | 1.04e+003 | 3.67e+003 |
| N | 9.00e+000 | 4.38e+001 | 2.08e+003 | 1.33e+002 | 9.94e+002 | 2.19e+003 |

The attributes with observable discrimination potential between the *N* class against the other quality classes are printed in bold

**Table 16** Logistic regression model for teams on wiki-pl dataset

| | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | −1.071e+01 | 8.254e-01 | −12.980 | <2e-16 *** |
| **Versatility** | 1.730e+00 | 2.565e-01 | 6.743 | 1.55e-11 *** |
| Mean productivity in article | −2.252e-04 | 2.461e-05 | −9.153 | <2e-16 *** |
| Mean total productivity | 8.505e-09 | 1.446e-08 | 0.588 | 0.556 |
| **Size of team** | −2.176e-03 | 1.169e-03 | −1.861 | 0.0627 . |
| Mean tenure in article | −1.492e-02 | 8.297e-04 | −17.989 | <2e-16 *** |
| Mean tenure in wikipedia | 1.116e-04 | 9.325e-05 | 1.196 | 0.232 |
| **sd productivity in art** | 5.824e-05 | 5.636e-06 | 10.334 | <2e-16 *** |
| sd total productivity | −9.579e-08 | 1.482e-08 | −6.465 | 1.01e-10 *** |
| **sd tenure in article** | 8.797e-03 | 3.633e-04 | 24.215 | <2e-16 *** |
| sd tenure in Wikipedia | −5.259e-04 | 1.291e-04 | −4.074 | 4.63e-05 *** |
| Length | 5.202e-05 | 1.375e-06 | 37.823 | <2e-16 *** |
| Age | −4.449e-04 | 4.221e-05 | −10.540 | <2e-16 *** |

Signif. codes: p<0 '***', p<0.001 '**', p<0.01 '*', p<0.05 '.', p<0.1 ' '

We simply define the *team* assigned to an article as a group of editors who contributed to this article.

## 6.1 Attributes of teams

In this section we introduce and compute several attributes for editors and teams that will be used in our statistical analyses.

**Table 17** Logistic regression model for teams on wiki-de dataset

| | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | −1.167e+01 | 6.628e-01 | −17.614 | < 2e-16 *** |
| **Versatility** | 4.612e-01 | 2.315e-01 | 1.992 | 0.04632 * |
| Mean productivity in article | −1.950e-04 | 1.927e-05 | −10.120 | < 2e-16 *** |
| Mean total productivity | −1.869e-07 | 1.323e-08 | −14.126 | < 2e-16 *** |
| **Size of team** | 2.379e-03 | 2.719e-04 | 8.750 | < 2e-16 |
| Mean tenure in article | −1.741e-02 | 8.874e-04 | −19.620 | < 2e-16 *** |
| Mean tenure in Wikipedia | 1.499e-03 | 9.026e-05 | 16.602 | < 2e-16 *** |
| sd productivity in art | 3.170e-05 | 3.262e-06 | 9.718 | < 2e-16 *** |
| **sd total productivity** | 7.595e-08 | 4.947e-09 | 15.353 | < 2e-16*** |
| **sd tenure in article** | 7.421e-03 | 3.126e-04 | 23.737 | < 2e-16 *** |
| sd tenure in Wikipedia | −4.687e-04 | 1.470e-04 | −3.188 | 0.00143 ** |
| Length | 3.939e-05 | 7.311e-07 | 53.884 | < 2e-16 *** |
| Age | 5.340e-04 | 3.112e-05 | 17.162 | < 2e-16 *** |

Signif. codes: p<0 '***', p<0.001 '**', p<0.01 '*', p<0.05 '.', p<0.1 ' '

**Table 18** Evaluation measures on testing data for teams on wiki-pl and wiki-de datasets

| Measure | Logistic regression teams wiki-pl dataset | Logistic regression teams wiki-de dataset | Random forest model wiki-pl dataset | Random forest wiki-de dataset |
|---|---|---|---|---|
| Precision | 25.34% | 38.21% | **66.91**% | **58.93**% |
| Recall | 4.46% | 7.65% | 7.41% | 20.27% |
| Accuracy | 99.71% | 99.57% | 99.74% | 99.25% |
| F-measure | 7.58% | 12.75% | 13.34% | 30.17% |

In particular, we consider the *tenure* of an editor on Wikipedia in the article measured as the number of days spent on editing Wikipedia articles.

In total, in this section we consider 10 team attributes that will serve as explanatory variables in our models and analyses. The attributes are presented in Table 13. In this table some attributes are based on standard deviation, i.e. they may be also viewed as representing team diversity measures. We present in bold all diversity-related attributes in the Table. For general description of diversity measures that we use, we refer the reader to Sections 2.1 and 2.2.

In this part we also utilise the division into the same quality classes as before, however there is no article marked as good and featured at once in any of our datasets. Therefore it is irrelevant to consider the (G∩F) quality class in the context of a team assigned to an article. In logistic regression and prediction experiments we treat the class G∪$F$ as the "high quality" label (C=1), and normal (N) as the "normal quality" label (C=0) (similarly as for single editors).

The order of the coming sections and experiments concerning teams is generally analogous to the one concerning editors with some necessary adaptations.

## 6.2 Preliminary exploratory data analysis for teams

At the beginning of experiments we did an exploratory analysis of the mentioned team attributes (Table 13) and their relationship with work quality and found some of them as promising for discriminating quality classes. The results are presented in Tables 14 and 15 (where "sd" stands for "standard deviation"). Results for wiki-pl and wiki-de datasets signal similar relationships. For each dataset, we print in bold the group of all attributes that observably discriminate the normal (N) quality class versus higher quality classes. Out of 10 attributes only 5 in wiki-pl and 7 in wiki-de belong to this group. Interestingly, versatility does not belong to these groups, but other diversity-related attributes are highly present. As was with the case of editors, more sophisticated analysis will later demonstrate that team versatility actually is among the factors that are positively related with work quality.

More precisely, Tables 14 and 15 demonstrate that versatility, mean total productivity, standard deviation total productivity, mean tenure in Wikipedia and standard deviation tenure in Wikipedia seem to be indifferent between four group qualities. Versatility is just slightly higher for better quality articles than for normal ones. Total productivity of editors in teams seems to have no significant relationship with quality of articles. The most considerable differences are observed for the following attributes: mean productivity in article, standard deviation productivity in article, team size, mean tenure in article and standard deviation in article. It seems that productivity and tenure and their diversity have stronger relationship with quality, when measured in article than in the whole Wikipedia. It doesn't

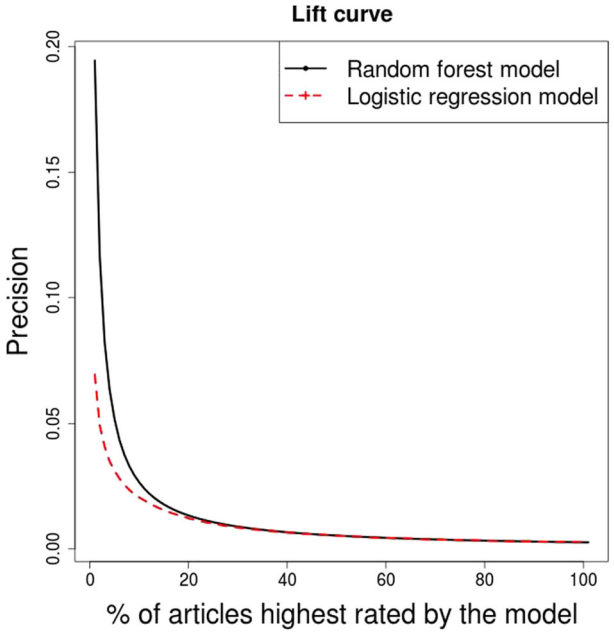


**Fig. 9** Lift curve for wiki-pl dataset models

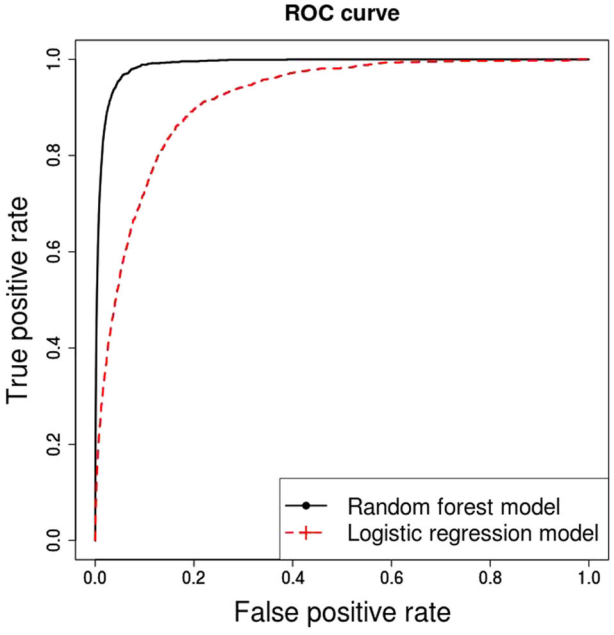


**Fig. 10** ROC curve for wiki-pl dataset models

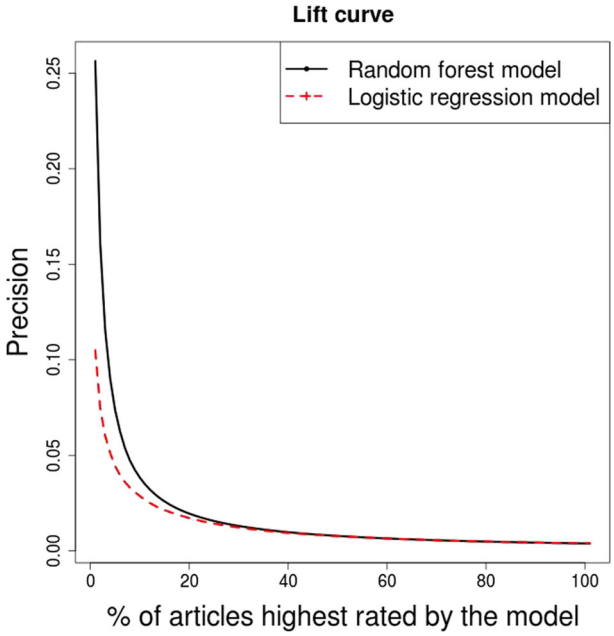


**Fig. 11** Lift curve for wiki-de dataset models

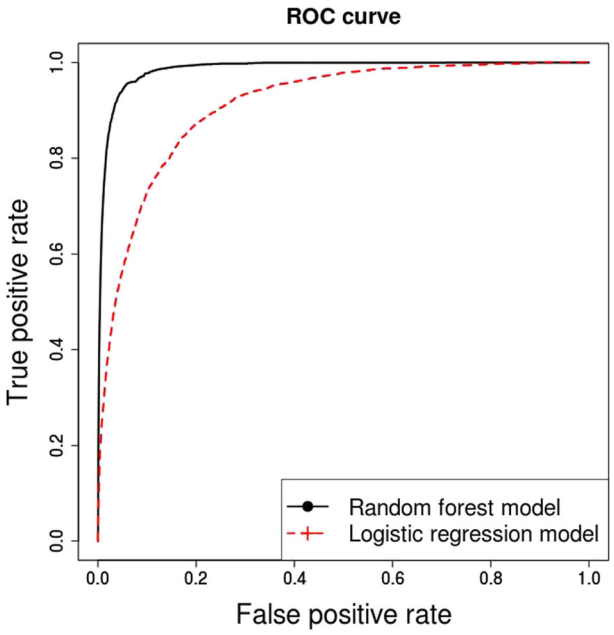


**Fig. 12** ROC curve for wiki-de dataset models

matter how much work was done by editors in all articles but only productivity in particular one article has relationship with it's quality. These results might indicate that "new" and "old" editors in an article through exchanging their experience create articles of better quality.

### 6.3 Logistic regression analysis

In this Section we fit a logistic regression model to the data using all 10 attributes described in Table 13 and $G \cup F$ as the target attribute.

Tables 16 and 17 show coefficient estimates, standard errors, Wald statistics (z-values) and their corresponding p-values for models.

Observed p-values demonstrate that almost all variables are statistically significant (assuming significance level 0.05), except mean total productivity for wiki-pl dataset and standard deviation tenure in Wikipedia for wiki-de dataset.

We highlight some more interesting observations in the tables by using bold print. Interestingly, in both datasets versatility has the absolutely highest positive coefficient of influence on quality, however it is statistically less significant than most of the other attributes. On the other hand, out of the three statistically strongest attributes (highest z-values) the majority (two) represent diversity-related attributes (standard deviations of productivity in article and of tenure or total productivity, depending on the dataset). In both datasets the remaining statistically strong (high z-value) attribute is team size, that is intuitionally obvious (large team likely improves the article).

### 6.4 Experiments with quality prediction for teams

In this Section we present experiments with prediction models concerning teams.

We used the aggregated data from the previous Section, split into training (50 % observations) and testing (50 % observations) datasets, and built logistic regression and Random Forest models (Liaw and Wiener 2002). Similarly as in case of editors, a response variable can take two values $C = 0$, which represents normal quality articles or $C = 1$, for both "higher" quality (G∪F). In other words we want to predict the probability of being a G∪F article produced by a team over the normal quality article.

As in the case of experiments for editors, we would like to verify how accurately it is possible to predict the quality of the article based on the features describing teams. Because the number of features is larger than for experiments with editors, instead of a single classification tree, we applied the Random Forest model, whose performance is usually superior to decision tree, and used the implementation available in the `R` package (`RandomForest`).

Table 18 shows evaluation measures for teams. The results indicate much lower prediction performance of logistic regression compared to the similar experiments concerning single editors. In general, Random Forest performs much better than logistic regression, the precision measure of Random Forest for both datasets is quite high here. It is larger for the wiki-pl data set (66.91 %) than for the wiki-de data set (58.93 %). Also other measures differ between the datasets. Thus, this experiment shows a different outcome than the corresponding experiment concerning single editors.

### 6.5 Lift and ROC curves

To further examine the prediction models we computed the Lift and ROC curves for our team quality prediction models (Figs. 9, 10, 11 and 12).

Figures 9 and 11 show Lift curves for teams of wiki-pl and wiki-de dataset. Figures 10 and 12 show the corresponding ROC curves. The results are significantly above the baseline. Note that, Random Forest outperforms logistic regression for both datasets. The ROC curves indicate that the Random Forest model performs better here than in the case of the experiments with single editors.

### 6.6 Importance of diversity measures in quality prediction

To additionally verify our hypothesis for teams, we assess the relevance of variables by using some variable importance measures available in the Random Forest model (Breiman 2001). The first measure (Imp1) is based on prediction error. Namely, for each tree, the prediction error on the out-of-bag portion of the data (data not used to build the model) is computed. Then the same is done after permuting the values of the given attribute (this makes the attribute irrelevant). The difference between the two are then averaged over all trees, and normalized by the standard deviation of the differences. The second measure (Imp2) pertains to average decrease of node homogeneity. Algorithms for constructing decision trees usually work top-down, by choosing an attribute at each step that best splits the set of observations. The quality of the split is measured using the decrease of node homogeneity, .e.g the difference between the class homogeneity in parent node and the child nodes. The class homogeneity is measured using entropy or Gini impurity measure. Large decrease indicates that the attribute is relevant. The average decrease of node homogeneity is taken over all splitting nodes and over all trees used to construct an ensemble classifier. Generally, the higher the value of the importance measures the stronger relationship with the predicted attribute (article quality).

Tables 19 and 20 show the results for wiki-pl and wiki-de dataset. For each dataset and importance measure we print in bold the attribute with the highest importance value.

All of the "winning" attributes in this analysis represent diversity-related attributes. Interestingly, for both datasets the winners are the same: versatility for the Imp1 measure and "standard deviation of productivity in article" for the second importance measure.

For both datasets, the diversity-related attributes like versatility and standard deviations are among the most significant variables according to either of the importance measures (Imp1, Imp2).

**Table 19** Random Forest importance for wiki-pl dataset. Imp1 is based on the differences in prediction errors. Imp2 is based on the average decrease of node impurity (see the details in the text)

| | Imp1 | Imp2 |
|---|---|---|
| Versatility | 3.77e+001 | 1.13e+002 |
| Mean productivity in article | 1.52e+001 | 1.01e+002 |
| Mean total productivity | 3.61e+001 | 9.55e+001 |
| Size of team | 1.77e+001 | 8.02e+001 |
| Mean tenure in article | 5.07e+000 | 6.15e+001 |
| Mean tenure in Wikipedia | 2.27e+001 | 7.55e+001 |
| sd productivity in art. | 1.07e+001 | 1.23e+002 |
| sd total productivity | **4.89e+001** | 1.02e+002 |
| sd tenure in article | 4.77e+000 | 6.67e+001 |
| sd tenure in Wikipedia | 2.72e+001 | 8.85e+001 |
| Length | -7.89e+000 | **1.53e+002** |
| Age | 2.42e+001 | 8.46e+001 |

**Table 20** Random Forest importance for wiki-de dataset

| | Imp1 | Imp2 |
|---|---|---|
| Versatility | 2.58e+001 | 5.02e+001 |
| Mean productivity in article | 1.68e+001 | 6.10e+001 |
| Mean total productivity | 1.59e+001 | 3.55e+001 |
| Size of team | 1.37e+001 | 5.62e+001 |
| Mean tenure in article | 8.74e+000 | 3.80e+001 |
| Mean tenure in Wikipedia | **3.35e+001** | 7.43e+001 |
| sd productivity in art. | 1.21e+001 | **9.14e+001** |
| sd total productivity | 1.74e+001 | 3.72e+001 |
| sd tenure in article | 8.20e+000 | 3.56e+001 |
| sd tenure in Wikipedia | 1.17e+001 | 3.51e+001 |
| Length | 1.59e+001 | 1.34e+002 |
| Age | 1.09e+001 | 4.23e+001 |

This result is especially significant, since we consider 10 attributes including such seemingly "strong" ones as "size of team" or tenure of editors. Diversity-based attributes turn out to be superior to them in this experiment.

### 6.7 Summary of experimental results for teams

This Section summarizes the experiments concerning teams of editors. Here our aim was to verify how different properties of teams (see Table 13), including diversity measures, influence the quality of articles. In this case we use two statistical models: logistic regression and random forest (more sophisticated ensemble of decision trees, tailored to the situation of larger number of attributes). The evaluation measures (Precision, Recall, F-measure) are again very promising. They are much larger than for the baseline (random assignment of articles to classes describing quality). Random forest outperforms the logistic regression significantly (this is clearly seen on ROC curves). As the performance of random forest was superior, we also calculated attribute importance measures based on random forest to check which attributes are useful for prediction of quality. It turns out that versatility is the most significant attribute according to the first measure. The experiments clearly indicate that diversity-related attributes of teams are strongly connected with the quality of the articles.

## 7 Conclusions and future work

In this article we applied statistical analysis to verify our hypothesis of whether diversity of editors and teams plays an important role in work quality in an open-collaboration environment on the example of Wikipedia.

A series of experiments ranging from more basic exploratory analyses to more advanced techniques including machine learning prediction models executed on two datasets positively verify our hypothesis.

We reported many statistical signals that diversity seems to play an important positive role in high quality cooperative work in Wikipedia.

Interestingly, some of the reported experiments indicated that the considered diversity-related attributes such as interest diversity (versatility) or experience diversity in teams

(st. dev. of tenure or st. dev. of productivity in team) are more connected with the quality of work than such "obvious" attributes as the average experience of the team members or even size of the team.

These findings give interesting insights into the studies of virtual open-collaboration communities and, as we hope, may motivate further work aimed at deeper analysis of the role of diversity in this context.

Another possible outcome of the study presented in this article would be to provide some valuable foundations for developing an intelligent decision-support system for suggesting how to build a successful virtual team in open-collaboration environment in order to produce high-quality outcome. In particular, it would be interesting to study in a future work whether the controlled level of diversity *intentionally* introduced to the team improves the quality of its work.

We hope that this work would serve as one of the steps towards achieving such goals in future.

**Acknowledgments** The work was partially supported by the Polish National Science Centre grant 2012/05/B/ST6/03364.

The study is co-financed by the European Union under the European Social Fund. Project PO KL "Information technologies: Research and their interdisciplinary applications", Agreement UDA-POKL.04.01.01-00-051/10-00.

We would like to thank J.Szejda and D.Czerniawska for their contributions to the early stage of the work that eventually resulted in this article.



## References

Aggarwal, A.K. (2014). Decision making in diverse swift teams: an exploratory study. In *47th Hawaii international conference on system sciences, HICSS 2014* (pp. 278–288). Waikoloa.

Agrawal, R., Gollapudi, S., Halverson, A., & Ieong, S. (2009). Diversifying search results. In *Proceedings of the 2nd ACM international conference on web search and data mining, WSDM '09* (pp. 5–14). New York: ACM.

Baraniak, K., Sydow, M., Szejda, J., & Czerniawska, D. (2016). Studying the role of diversity in open collaboration network: experiments on wikipedia. In *Advances of network science (Proceedings of the NetSci-X 2016 conference), Lecture Notes in Computer Science, Chap 8*, Vol. 9564. Springer.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Monterey: Wadsworth and Brooks.

Brown, C.D., & Davis, H.T. (2006). Receiver operating characteristics curves and related decision measures: a tutorial. *Chemometrics and Intelligent Laboratory Systems*, *80*(1), 24–38.

Carbonell, J., & Goldstein, J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '98* (pp. 335–336). New York: ACM.

Chen, J., Ren, Y., & Riedl, J. (2010). The effects of diversity on group productivity and member withdrawal in online volunteer groups. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 821–830). ACM.

Cormen, T.H., Stein, C., Rivest, R.L., & Leiserson, C.E. (2001). *Introduction to algorithms*, 2nd edn: McGraw-Hill Higher Education.

Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., & Suri, S. (2008). Feedback effects between similarity and social influence in online communities. In *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '08* (pp. 160–168). New York: ACM.

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, *438*, 900–901.
Goffman, W. (1964). A searching procedure for information retrieval. *Information Storage and Retrieval*, *2*(2), 73–78.
Hosmer, D.W., Lemeshow, S., & Sturdivant, R.X. (2013). *Applied logistic regression*. New York: Wiley.
Kittur, A., & Kraut, R.E. (2008). Harnessing the wisdom of crowds in wikipedia: Quality through coordination. In *Proceedings of the 2008 ACM conference on computer supported cooperative work, CSCW '08* (pp. 37–46). New York: ACM.
Langlois, R.N., & Garzarelli, G. (2008). Of hackers and hairdressers: modularity and the organizational economics of open-source collaboration. *Industry and Innovation*, *15*(2), 125–143.
Liaw, A., & Wiener, M. (2002). Classification and Regression by random. *Forest R News*, *2*, 18–22.
López, C.A., & Butler, B.S. (2013). Consequences of content diversity for online public spaces for local communities. In *Proceedings of the 2013 conference on Computer supported cooperative work* (pp. 673–682). ACM.
Parnas, D.L. (1972). On the criteria to be used in decomposing systems into modules. *Communications of the ACM*, *15*(12), 1053–1058.
R Core Team (2013). R: A Language and Environment for Statistical Computing. Technical report, R Foundation for Statistical Computing.
Ren, Y., Chen, J., & Riedl, J. (2015). The impact and evolution of group diversity in online open collaboration. *Management Science*.
Sanchez, R., & Mahoney, J.T. (1996). Modularity, flexibility, and knowledge management in product and organization design. *Strategic Management Journal*, *17*(S2), 63–76.
Shannon, C.E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423.
Strzezek, A., Trammer, L., & Sydow, M. (2015). Divergene: experiments on controlling population diversity in genetic algorithm with a dispersion operator. In *Proceedings of the 2015 federated conference on computer science and information systems, annals of computer science and information systems*, (Vol. 5 pp. 155–162).
Sydow, M., Pikuła, M., & Schenkel, R. (2013). The notion of diversity in graphical entity summarisation on semantic knowledge graphs. *Journal of Intelligent Information Systems*, *41*(2), 109–149.
Therneau, T., Atkinson, B., & Ripley, B. (2015). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-10.
Vasilescu, B., Posnett, D., Ray, B., van den Brand, M.G., Serebrenik, A., Devanbu, P., & Filkov, V. (2015). Gender and tenure diversity in github teams. *CHI. ACM*.
Vee, E., Srivastava, U., Shanmugasundaram, J., Bhat, P., & Yahia, S.A. (2008). Efficient computation of diverse query results. In *IEEE 24th international conference on data engineering, 2008. ICDE 2008* (pp. 228–236). IEEE.
Wilkinson, D.M., & Huberman, B.A. (2007). Cooperation and quality in wikipedia. In *Proceedings of the 2007 international symposium on wikis, WikiSym '07* (pp. 157–164). New York: ACM.