Universität Ulm

Fakultät für Mathematik und
Wirtschaftswissenschaften

A probability-based approach for measuring the currency of Wiki articles

Masterthesis

In Wirtschaftsmathematik

vorgelegt von
Moestue, Lars
am 26.02.2021

Gutachter

Prof. Dr. Mathias Klier
Prof. Dr. Steffen Zimmermann

Table of contents

List of figures

List of tables

List of abbreviations

AUC	Area under the curve
Cdf	Cumulative distribution function
IS	Information system
LSTM	Long short-term memory
NLP	Natural language processing
Q-Q-plot	Quantile-Quantile plot
ROC	Receiver operating characteristic

Acknowledgements

First and foremost, I would like to thank Prof. Mathias Klier, Andreas Obermeier, and Torben Widman for giving me the opportunity to work on this interesting project and for all the fruitful and interesting discussion we had together.

Secondly, I would like to thank Prof. Steffen Zimmermann for agreeing to be the second reviewer of this thesis.

Lastly, I would like to thank Leana, Lena, and Nora for carefully proofreading this thesis and calling my attention to several linguistic and grammatical errors.

1 Introduction

Wikis nowadays are the central and best way to share information across many different people (McAfee 2006). Wikipedia, the most famous example for a Wiki, is considered to be the most comprehensive knowledge repository in human history (Dang and Ignat 2017). The importance of Wikipedia is unbroken as alone the five most important Wikis of Wikipedia (English, German, Spanish, French, and Russian) had over 137 billion pageviews in 2020[1] and therefore Wikipedia is one of the most important websites on the internet[2]. Shortly after its launch, it already became a central information source for organizations as well as for individuals (Blumenstock 2008a). Based on this success Wikis became the main platform for knowledge management and transfer in companies as well (Pfaff and Hasan 2006; Richter et al. 2013). As a matter of fact, different studies showed that Enterprise Wikis are used in a wide range of industry sectors like Microelectronics, Engineering services, IT services (Stocker and Tochtermann 2011), software development (Trkman and Trkman 2009), aviation, oil companies, and consumer goods corporation (Pei Lyn Grace 2009). Beyond knowledge saving Wikis are used for a wide variety of different tasks such as group work (Morgan et al. 2013), educational settings (Fominykh et al. 2016 Heidrich et al. 2015; Matschke et al. 2013), and even collaborative law reforms (Aitamurto et al. 2017). However, a great challenge for Wikis is to hold a high-quality standard as they are open projects with almost no constraints on who can edit articles (Klobas 2006; Peacock et al. 2007). Moreover, Kiniti and Standing (2013) and Klobas (2006) describe sufficient data quality as a key factor of the success of a Wiki, as it correlates with the trust of the users.

The concept of data quality, sometimes referred to as information quality can be used to characterize mismatches between the true state of the world and the view of the world provided by an information system (IS) (Orr 1998; Parssian et al. 2004). It is a multidimensional construct (Laranjeiro et al. 2015; Lee et al. 2002; Redman 1997) comprising several values such as accuracy, completeness, currency, and consistency (Wang et al. 1995b). Recent studies (Spruit and van der Linden 2019; QAS 2013) have revealed that one of the most common defects is outdated data. This is especially important in Wikis since the use and satisfaction of Wikis users is highly correlated to the currency of the Wiki (Bhatti et al. 2018). In enterprises, an up-to-date Wiki, therefore, leads to higher individual efficiency and capacity development (Bhatti et al. 2018; He and Yang 2016; Stefanovic et al. 2016). Manual approaches to ensure high data quality in Wikis require extensive use of curators, who in a

[1] https://pageviews.toolforge.org/siteviews/?platform=all-access&source=pageviews&agent=user&start=2020-01&end=2020-12&sites=en.wikipedia.org|de.wikipedia.org|fr.wikipedia.org|ru.wikipedia.org|es.wikipedia.org

[2] https://www.alexa.com/topsites

cumbersome process constantly correct and update articles. This process is particularly expensive and maintaining a high data quality even seems impossible. An automated approach for assessing the currency in Wikis could lead to enormous cost savings. Despite their relevance, there is still a lack of well-founded and applicable data quality metrics to assess the currency of Wikis. In practice, articles often need to be updated after an event of high interest of the Wiki users such as the death of a person or the appointment of a new CEO in a company. Based on this idea we propose a probability metric to measure such events to evaluate if a certain article needs to be updated.

The remainder of this work is organized as follows. To ensure relevance and practical utility of the proposed metric the problem context is illustrated in Chapter 2. The metric is designed to calculate probabilities of notable events that need to result in updates in the corresponding Wiki article based on statistical outlier detection. To guarantee that the proposed approach is based on a solid theoretical foundation in Chapter 3 the necessary theory for the development of the event-based probability metric is outlined. This includes an overview of different outlier detection methods and data quality dimensions and existing metrics. Moreover, the evaluation methods of probability-based metrics are illustrated. On this basis Chapter 4, provides an overview of the related work to identify the body of knowledge and carve out the research gap to be addressed. Bringing together these lines in Chapter 5 a novel approach for measuring the currency of Wiki articles is proposed and derived. Furthermore, a detailed description of the practical calculation is given. Applying this approach on a data set of the English Wikipedia the practical applicability and benefit is demonstrated in Chapter 6 using the theoretical foundations from Chapter 3. Finally, in Chapter 7 the results are summarized, and the limitations of the approach and future research are discussed.

2 Problem Context

Currency expresses a time-related data quality dimension, however, the definitions in literature are less uniform. There are many different terms used such as currency, timeliness, up-to-date and temporal validity. Some contributions use the same term for a different concept, while others define similar concepts with different terms. Nelson et al. (2005) and Redman (1997) use currency to describe the degree to which a datum is up-to-date. An analogue approach is addressed by Cho and Garcia-Molina (2003) as well as Xiong et al. (2008). However, they refer to it as freshness of the data. Moreover, Cho and Garcia-Molina (2003) use the term up-to-date if a stored data equals its real-world counterpart. In contrast, Ballou et al. (1998) defines currency as the age of an attribute value and referring to timeliness as whether the recorded value is not out of date. Further, they call the time data remains valid shelf life. Other authors such as Batini and Scannapieco (2016) define currency in the context of how promptly data is updated, while timeliness expresses how current the data are at hand.

Heinrich and Klier (2015) use the term currency to express whether a value in an IS is still equivalent to the real-world at the instant of assessment. This is only a brief and superficial discussion of time-related data quality dimensions, but it illustrates that no widely accepted definition has prevailed yet. We will build primarily upon the definition of Redman (1997) and Heinrich and Klier (2015) and use the term up-to-date.

A Wiki article is up-to-date if all contained information is still the same as in the real world and if the article contains all relevant information from the recent past. Otherwise, an article is outdated. In the context of Wikis information in articles can only be outdated if there is a change to the corresponding entity in the real-world. Such changes in the real-world can only happen if there is an event in the real-world. An event in this approach is meant very widely. Examples for such events range from elections and a designation of a new CEO to a publication of new population numbers of a city or the relocation in another office of the HR department. All these examples would need to result in an update of a Wiki. Both the individual's Wikipedia page, who won the election, and the article about the city, which population data got republished, would need to be updated. In an enterprise Wiki, the position of the new CEO has to be changed, as well as the new bureau of the HR department so everybody knows where to locate them. Events like an election and the designation of new CEO draw the interest of many Wiki users. We define such an event that has the attention of a lot of Wiki users as notable event. Our approach aims to identify notable events in order to determine articles that need to be updated. The detection of notable events is very important and relevant as they need to result in an update in the corresponding Wiki article. If there is no update after a notable event the Wiki article would be outdated as the information of the notable event would be missing. For non-notable events such as the new bureau already existing approaches can be used to detect these events such as Ballou et al. (1998) and Heinrich and Klier (2015) for the case of regular updates such as population data or Pernici and Scannapieco (2003) in the case of known expiry dates, such as moving dates.

However, these approaches do not work for the detection of notable events since they happen in unregular time intervals and are mostly unforeseeable. As a consequence, approaches that work with known shelf lives or expiry dates do not work for detecting notable events. Nevertheless, notable events are highly relevant and need to be mentioned as many Wiki users are interested in these events. Thus, it is necessary to develop an approach to measure such notable events. In other contexts currency is frequently measured with probability-based approaches (Heinrich et al. 2009, Heinrich and Klier 2015, Wechsler and Even 2012, Zak and Even 2017, Zong et al. 2017). The advantage of probability-based approaches is that they can cope with uncertainty. Therefore, it is very useful in our context as well, as our goal is to identify notable events, which are highly unregular and unforeseeable. Moreover, only by definition, it fulfils many requirements for data quality metrics proposed by Heinrich et al. (2018b) such as the existences of a minimum and maximum, interval scaling, and interpretability. Especially the latter is an important advantage in comparison to other existing approaches such as Stvilia et al. (2005b), who measure currency as time since the last update. Furthermore, it could be used in decision calculus. For instance, it is possible to implement an automatic system that notifies a certain number of subscribers or administrators of an article based on the probability of an event. These subscribers and administrators then could check what kind of event happened and if the article is already updated. Recapitulatory it can be stated that a probability-based approach is most suitable to measure notable events, therefore the aim of this work to develop a metric to detect notable events based on probability theory.

3 Theoretical Foundations

For the development and evaluation of the in the previous chapter described and in chapter 5 proposed probability-based metric a solid theoretical foundation, presented in this chapter, is necessary. As the metric aims to detect notable events, which can be described as an outlier, in the first section different statistical outlier tests are presented and discussed. To show the relevance of the proposed metric in the second section different metrics for different data quality dimensions are presented and discussed with respect to the applicability in Wikis. The last section focuses on the evaluation of probability-based metrics to give the theoretical foundations necessary to evaluate the proposed metric in chapter 6.

3.1 Statistical outlier tests

The goal of this section is to find an outlier test based on a rigorous mathematical foundation, free of subjectiveness, that can be used to assess outliers automatically and make reliable statistical statements for data sets with a large number of observations. Therefore, the most important outlier tests (Hodge and Austin 2004; Walfish 2006) are presented.

In statistics, an outlier in a data set is a data point that differs significantly from other observations (Grubbs 1969, Maddala and Lahiri 1992). Outliers are often caused by variability in the measurement, experimental errors, or extraordinary events (Grubbs 1969). However, there is no rigid mathematical definition of what constitutes an outlier. In practice, there are several methods of outlier detection. These can be distinguished in two fields, graphical methods and model-based approaches.

3.1.1 Graphical methods

As the name suggests, graphical methods are based on plots. They require a manual assessment of an individual to decide if a certain point is an outlier. The remainder of the section presents the two most important graphical methods for outlier detection.

Box plots

Box plots were first introduced by the American mathematician John W. Tukey (1970, 1977). In this work, mainly the work of Wickham and Stryjewski (2010) is used to explain the idea of a box plot. Figure 1 shows an example of a box plot. The box in the middle represents the interquartile range, which is defined as the interval between the first quartile (25%-percentile) and the third quartile (75%-percentile). Thus, the interquartile range contains 25% of all values that are below the median and closest to it and 25% of all values that are greater than the median and closest to it. This means in total the interquartile range contains 50% of all values. The median is shown as a horizontal line inside the box (cf. the orange line in Figure 1). The range of the values outside the interquartile range (all values between the minimum and the first quartile and between the third quartile and the maximum) is presented as a straight vertical line, which is called whisker. The minimum and maximum are marked by a horizontal line at the end of the whisker. Indicators for an outlier could be extraordinary long whiskers or whiskers that have exceedingly different lengths.

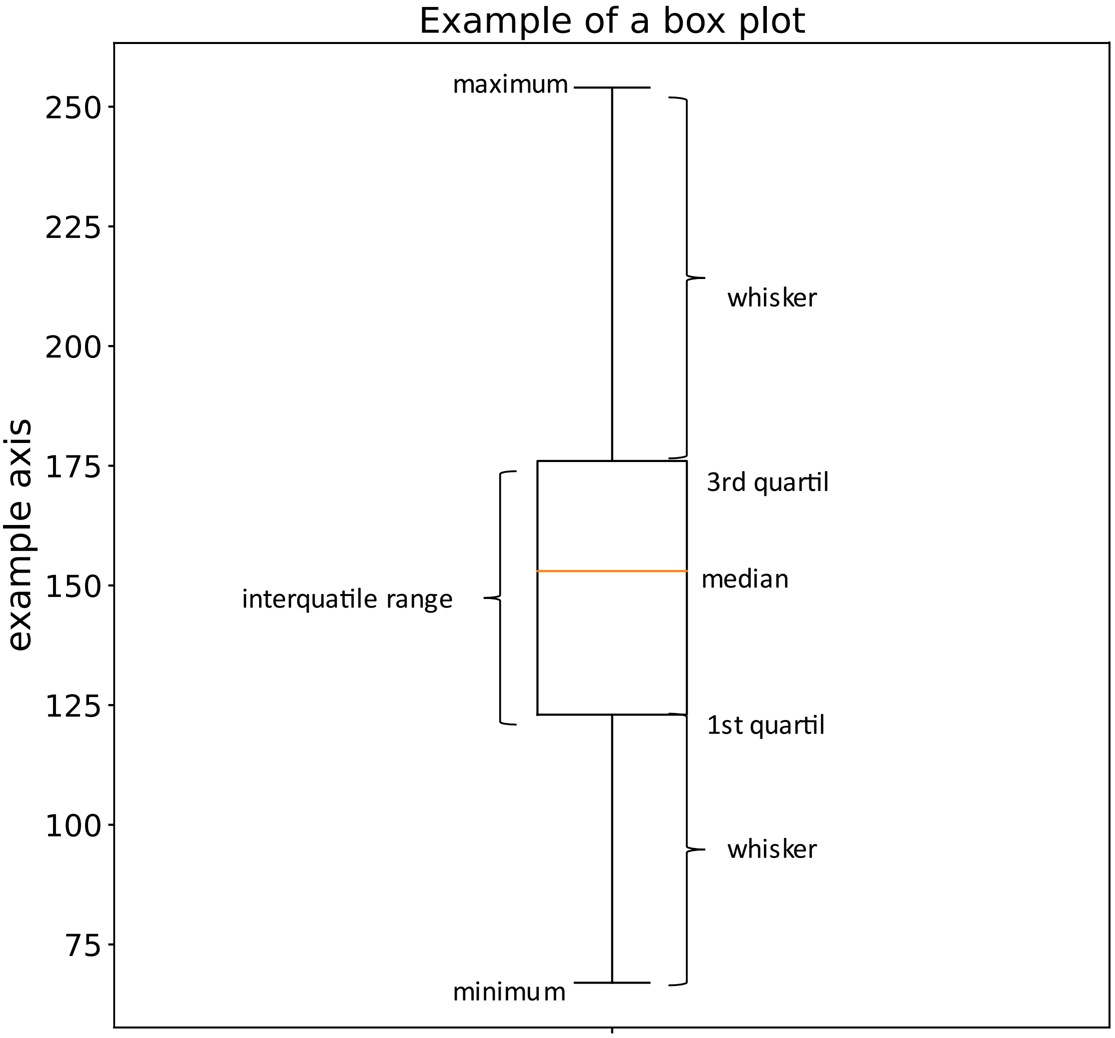

Figure 1: Example of a box plot

Nevertheless, without knowing the underlying distribution, it is a difficult task to detect an outlier only based on a box plot. Problems especially can occur with heavy-tailed distributions or skewed distributions. In these cases, long whiskers or whiskers with different lengths can be misleading as they are expected to occur with such distributions. Thus, if the underlying distribution of the data set is known, an approach that takes these distributions into account is more suitable. A very prominent example of a graphical method for outlier detection considering a distribution is Q-Q plots presented next.

Quantile-Quantile plot (Q-Q-Plots)

Q-Q-plots are usually used in statistics to verify if a data sample follows an assumed underlying distribution. However, they can also be applied to analyse whether there is an outlier in the data set. To obtain a Q-Q-plot, knowing or having an assumption of the theoretical distribution of the data set is necessary. The approach is explained following the work of Thode (2002).

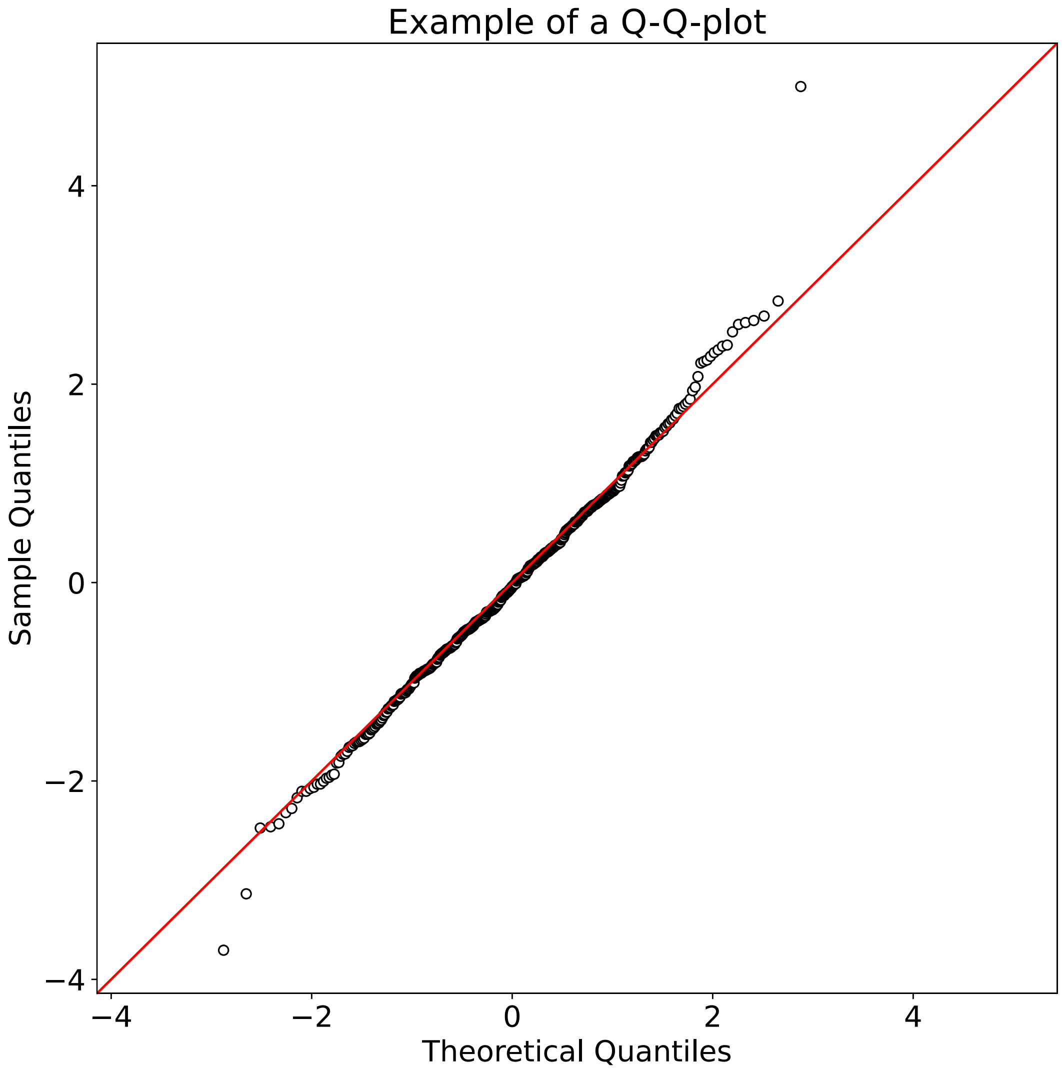

Figure 2: Example of a Q-Q-plot

In a Q-Q-plot, the sample quantiles are plotted against the expected theoretical quantiles. The sample quantiles can be easily attained by ordering the data in ascending order. In a data set with n data points the $\frac{i}{n}$-quantile ($i \in \{1, \ldots, n\}$) would be the i^{th} smallest data point. This quantile is then usually plotted against the $\frac{i-0.5}{n}$-quantile. This small adjustment is applied to avoid problems with infinite values in the case of unbounded distributions for $i = n$ (as the 100%-quantile for unbounded distributions is infinity). If the sample distribution is the same as the assumed theoretical distribution the plot will closely follow the angle bisector ($y = x$). Outliers can be determined if the plot is close to the angle bisector except for a few points in the bottom left corner, which are far below the expected line or in the upper right corner far above the expected line. In such cases, the data set overall follow the expected distribution except for the minimum resp. the maximum. The minimum would be much smaller than expected as the observed quantile is much smaller than the expected quantile or the maximum would be much greater than expected since the observed quantile would be much greater than the expected quantile. In Figure 2 for instance, the point in the upper right corner is probably an outlier as it is far above the expected line. Moreover, the two minimum points in the bottom left corner are below the expected line and thus could be a possible outlier. Nevertheless, it is feasible as well that these points occur just randomly. Although Q-Q-plots are a more advanced method than box plots for detecting outliers, it remains a very subjective exercise to rely on graphical methods only. Therefore, in the next subsection more well-founded approaches are presented.

3.1.2 Model-based approaches

Model-based approaches assume an underlying distribution of the data set. In all methods presented in this subsection, except Peirce's criterion, this assumed underlying distribution is a normal distribution. From a statistical point of view, model-based approaches are simple hypothesis tests (for an introduction in hypothesis tests cf. Wilcox (2011)). The null hypothesis H_0 is that the data set contains no outliers, while the alternative hypothesis H_1 is that there is an outlier in the data. Thus, these approaches are, if mathematically correctly derived, statistically well-founded and free of subjectivity (if there is a predefined significance level α). In the following, the most important model-based approaches are introduced and analysed carefully. Therefore, $X = \{X_1, \ldots, X_n\}$ is assumed to be a data set of n observations. The goal is to determine if the maximum or minimum value is an outlier at a significance level $\alpha \in (0, 1)$. Frequently used terms are the sample mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and the corrected sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Grubbs outlier test

This widely used outlier test is named after the American statistician Frank E. Grubbs. He is widely believed to be the first, who presented this approach (Adikaram, K. K. L. B. et al. 2015; Aslam 2020; Jain 2010; Wilrich 2013). However, like often in the history of mathematics (Grattan-Guiness and Ledermann 1994; Merzbach and Boyer 2011; Samelson 2001; Stigler 1980), this honours the wrong person. During the time of his dissertation, Grubbs worked on outlier tests. In 1950 he published a work on an outlier test developed by him, which nowadays normally is cited if the Grubbs test is referenced (Grubbs 1950). The proposed outlier test of Grubbs in this paper uses a different test statistic although it is equivalent to the nowadays used Grubbs test. However, it was the Canadian entomologist William R. Thompson, who was the first one to propose the test as it is used today and proving its rigour (Thompson 1935). Consequently, it was named Thompson's criterion in the following years (Pearson and Sekar 1936). Grubbs even referenced the work of Thompson in his work of 1950. It is not entirely clear at which time the work of Thompson was almost forgotten, but it was probably around 1970. In 1969 Grubbs published another paper (Grubbs 1969). This work was compared to the other papers rather easy to read and understand as it does not contain any mathematical proofs but only instruction on how to detect outliers. In this work, he recommends Thompson's test statistic (with a slightly and trivial adjustment), but without referencing it to Thompson. Instead, he references his work of 1950. This led to the current state, where the method is mostly known as Grubbs test with only a few calling it Thompson's tau (Shen et al. 2017) as it probably should be called.

The approach is presented following the work of Thompson (1935), Pearson and Sekar (1936), Grubbs (1969), and Stefansky (1972). Thompson was able to show that

$$\widetilde{G_\iota} = \frac{X_i - \bar{X}}{\tilde{s}},$$

where $\tilde{s}$ is the uncorrected sample standard deviation, which can be obtained by multiplying the corrected sample standard deviation s with $\sqrt{\frac{n-1}{n}}$, follows the same distribution as

$$t_{n-2}\sqrt{\frac{n-1}{n-2+t_{n-2}^2}},$$

with t_{n-2} being a studentized t-distribution with $n-2$ degrees of freedom.

Thus, with probability $1 - \alpha$ all observations $\{\widetilde{G_1}, \dots, \widetilde{G_n}\}$ are within the interval

$$\left[t_{n-2,1-\frac{\alpha}{2n}}\sqrt{\frac{n-1}{n-2+t_{n-2}^2}}\,, t_{n-2,\frac{\alpha}{2n}}\sqrt{\frac{n-1}{n-2+t_{n-2}^2}}\right].$$

Therefore, defining the following test statistic

$$\tilde{G} = \max_{i\in\{1,\dots,n\}} \frac{|X_i - \bar{X}|}{\tilde{s}},$$

the hypothesis of no outlier (meaning there is an outlier in the data) is rejected if it exceeds the critical value

$$\tilde{G} > t_{n-2,\frac{\alpha}{2n}}\sqrt{\frac{n-1}{n-2+t_{n-2}^2}}.$$

In applications and literature nowadays, a slightly modified test statistic and critical value proposed by Grubbs (1969) is used. Instead of using $\tilde{G}$, the alternative test statistic G defined as

$$G = \max_{i\in\{1,\dots,n\}} \frac{|X_i - \bar{X}|}{s},$$

with the corrected sample standard deviation instead of the uncorrected one, is applied. Therefore, the new critical value can be obtained by simply rearranging the critical value as

$$\frac{n-1}{\sqrt{n}}\sqrt{\frac{t^2_{\frac{\alpha}{2n},n-2}}{n-2-t^2_{\frac{\alpha}{2n},n-2}}}.$$

This version of the Grubbs test can identify outliers on both sides (minimum and maximum). However, there is a one-sided version of the Grubbs test, too. This is especially useful if only very high or very low outliers are of interest. For these cases only two small adjustments are necessary. Obviously, the test statistic only considers one side and therefore is

$$G = \max_{i\in\{1,\dots,n\}} \frac{X_i - \bar{X}}{s}, \text{resp.} \max_{i\in\{1,\dots,n\}} \frac{\bar{X} - X_i}{s}.$$

The first formula is used if it is tested whether the maximum value is an outlier, while the second formula is used to test if the minimum is an outlier. Moreover, the critical value is adjusted by using $t^2_{\frac{\alpha}{n},n-2}$ instead of $t^2_{\frac{\alpha}{2n},n-2}$.

Chavuenet's criterion

First introduced by William Chavuenet in 1871 (Chavuenet 1871) it is one of the most widely used criteria for outlier rejection nowadays (Ross 2003). The idea of this approach is explained based on the work of Taylor (1997). It is used to decide whether the minimum resp. the maximum observation $X_o \in \{X_1, \dots, X_n\}$ can be expected to occur. Based on the assumption of a normal distribution $\mathcal{N}(\bar{X}, s)$ the probability $p = P(X > X_o)$ resp. $p = P(X < X_o)$ is calculated.

This can be done either using numerical algorithms or by transforming X_o into a standard normal distributed variable by $\widetilde{X_0} = \frac{X_o - \bar{X}}{s}$ and using a standard normal cumulative distribution function (cdf) table to look up p. Such a table can be found in almost all fundamental statistics books such as Feller (2008). X_o is determined to be an outlier if $p < \frac{1}{2n}$. This criterion however is criticized in literature as there is no existing theoretical justification. Instead, the critical value $\frac{1}{2n}$ is chosen arbitrarily (Ross 2003). Moreover, Limb et al. (2017) showed in a Monte-Carlo simulation that using this criterion to reject outliers often results in worse estimates, especially for small n.

Dixon's Q-test

This outlier test was proposed by the American mathematicians Robert B. Dean and Wilfrid J. Dixon in 1951 and is especially useful for small data sets (Dean and Dixon 1951). The test is relatively simple. Without loss of generality, it is assumed that the data set $\{X_1, \ldots, X_n\}$ is ordered with X_1 being the smallest data value and X_n being the largest data value. This can be achieved by rearranging the values if necessary. Moreover, for simplicity it is assumed that it is tested if the maximum value X_n is an outlier. The gap is defined as difference between the maximum value and the next smaller value (gap $= X_n - X_{n-1}$) and the range as difference between the maximum value and the minimum value (range = $X_n - X_1$). Using these definitions, the test statistic Q is defined as

$$Q = \frac{gap}{range}.$$

To decide whether X_n is an outlier, Q needs to be compared to Q_α from tables which can be found in Dean and Dixon (1951) or Verma and Quiroz-Ruiz (2006). If $Q > Q_\alpha$, the hypothesis of no outlier is rejected and therefore X_n is an outlier. However, this approach of outlier detection has two major demerits. Firstly it is only suitable for small data sets with $n < 30$ (Dean and Dixon 1951), strongly limiting the application of this approach since nowadays most data sets containing larger numbers than only 30 data points. Secondly, the computation of the corresponding critical values Q_α is a time-consuming process as the formulas are very complicated. Moreover, corresponding tables (Verma and Quiroz-Ruiz 2006) only include values for several significance levels α, making it difficult to use this test for significance levels besides the standard levels.

Peirce's criterion

Peirce's criterion has a long history but is only used recently on a larger scale. It was first presented by American mathematician Benjamin Peirce in 1852 (Peirce 1852). However, the method requires multiple estimations, which need to be looked up in tables. Some of these tables were only computed and presented by Gould (1855) in a letter to the American physicist Alexander Dallas Bache. For this reason, this criterion was unpopular since it was much more complicated to compute in comparison to other approaches that are presented above. In the early 2000s, it seemed to have fallen into complete oblivion, until the American engineer Stephen M. Ross brought it back to the scientific community (Ross 2003). This time, with the advent of computers and fast numerical algorithms, it was much easier and less cumbersome to apply this approach. In the following, the approach is described in detail based on the work of Ross (2003).

Peirce's criterion is used to find outliers in observation data that are obtained to perform a regression. This means it is assumed that the observed data is the dependent variable (response variable) of a linear regression. The most elementary case would be pairs of observations in order to perform a simple linear regression. In addition, the approach also works on higher dimensional regression approaches. In the following, the number of explanatory variables (model unknowns) is called m ($m = 2$ would be a simple linear regression). In comparison to other approaches, Peirce's criterion can be used to identify multiple outliers at once. To present the approach n' is defined as the number of assumed outliers. The basic idea is to use the test statistic $T = \max_{\mathrm{i} \in \{1,\dots,\mathrm{n}\}} |X_i - \bar{X}|$. The hypothesis of no outlier is rejected if $T > sR$. To obtain the parameter R it is necessary to solve a system of non-linear equations:

$$\lambda = \left(\frac{Q^n}{r^{n'}}\right)^{\frac{1}{n-n'}}$$

$$R^2 = 1 + \frac{\mathrm{n} - \mathrm{n}' - \mathrm{m}}{\mathrm{n}}(1 - \lambda^2)$$

$$r = e^{\frac{R^2-1}{2}} \frac{2}{\sqrt{\pi}} \int_{\sqrt{\frac{R^2}{2}}}^{\infty} e^{-t^2} dt,$$

with

$$Q = \frac{n'^{\frac{n'}{n}} (n - n')^{\frac{n-n'}{n}}}{n}.$$

Nowadays numerical methods are capable of solving these kinds of systems of non-linear equations by iterating over all equations until the values converge (for a detailed description of such methods cf. Dennis and Schnabel (1996)). The derivation of these functions however is complicated and partly based on estimations. Furthermore, at least some of the assumptions on which this approach is based, contradict fundamental axioms of probability theory as shown by Dardis (2004), who proved that some of the probabilities that occur in the original paper can be greater than one. Moreover, several estimations are hardly justifiable. For instance, Peirce proposed a formula for the probability of the deviation of a certain observation from the mean. However, using this formula one gets comparable high probabilities for deviations from the mean if the expected deviation from the mean is smaller and vice versa. Obviously, this is a serious deficiency of this approach and contradicts the statements that “Peirce’s criterion is a rigorous method based on probability theory” (Ross 2003). Besides the shown mathematical weaknesses, Limb et al. (2017) showed that using this approach leads to similar problems as Chavuenet’s criterion as rejecting outlier based on this approach leads to worse estimations in a Monte-Carlo simulation.

3.1.3 Summary of the outlier tests

As stated at the beginning of this section the goal is to find an outlier test, which is based on a rigorous mathematical foundation and free of subjectiveness. In addition to that, it is required that it can be used to assess outliers automatically and make reliable statistical statements for data sets with a large number of observations. Table 1 shows an overview of the different approaches and which requirements each approach fulfils. Obviously, graphical methods such as box plots and Q-Q plots cannot fulfil all of the stated requirements. The decision if there is an outlier in a data set with graphical approaches is always made by a person. For that reason, this makes the assessment of outliers a very subjective matter. On the same data set different individuals will have different opinions whether there is an outlier in the data set. Even if there would be a standard to which all decision-makers would agree, it still would be necessary that a person makes the decision based on a plot. This violates the requirement that the method assesses the outlier detection automatically. Thus, these graphical approaches are disregarded, and only model-based approaches are considered. In comparison, all model-based approaches fulfil these two requirements. They all have comprehensible and fixed rules to determine if a data set contains an outlier. This makes these approaches free of subjectiveness. Furthermore, all approaches can evidently be automated in numerical computations. However, some of them are not fulfiling the other requirements. Dixon’s Q-test can only be applied to small data. For large data sets, the test is not able to make reliable statistical statements. Other approaches have weaknesses in the theoretical foundation.

	No subjectiveness	Can be used automatic	Rigour foundation	Work in large data sets
Box plots	×	×	✓	✓
Q-Q plots	×	×	✓	✓
Dixon’s Q-test	✓	✓	✓	×
Chavuenet’s criterion	✓	✓	×	✓
Peirce’s criterion	✓	✓	×	✓
Grubbs outlier test	✓	✓	✓	✓

Table 1: Overview of the fulfilment of the requirements in the different outlier tests

In Chavuenet’s criterion, the critical value is determined without any rigorous mathematics but chosen arbitrarily. The justification of Peirce’s criterion is partly contentious, partly inconsistent with fundamental axioms of probability theory. Furthermore, both of these approaches if used for outlier rejection frequently make Monte-Carlo estimations worse. Thus, both approaches do not have a rigorous mathematical foundation and therefore cannot be used to make reliable statistical statements. As a result, all three methods are neglected from further consideration. The only approach fulfiling all four requirements is the Grubbs outlier test (even if it should probably have another name) as it has neither any flaws in the theoretical foundation nor a limitation on the number of observations. This makes the Grubbs outlier test the perfect fit for our approach for measuring currency, which will be introduced in chapter 5.

3.2 Data quality dimensions and data quality metrics

Data quality is a highly researched topic with a variety of definitions in the literature (Batini et al. 2014, Laranjeiro et al. 2015, Ge and Helfert 2007). A majority defines data quality either as the fitness of use of the data (Sebastian-Coleman 2015, Strong et al. 1997) or as the degree to which the view of the world provided by an IS coincides with the true state of the world (Parssian et al. 2004; Batini and Scannapieco 2016). This work relies on the second definition (except for accessibility as foundation for all other data quality dimensions). Data quality in this context is seen as a multidimensional construct comprising several dimensions.

Data quality dimension	Short description (based on Laranjeiro et al. (2015))
Accessibility	The degree to which data can be assessed in a specific context.
Accuracy	The degree to which data attribute value correctly represents the real-world value counterpart.
Completeness	The degree to which all expected attributes are existing in the IS.
Consistency	The degree to which the data and information being compatible with other similar information objects.
Currency	The degree to which the data in the IS is up-to-date.

Table 2: Overview of the most important data quality dimensions

Table 2 offers an overview of the most important data quality dimensions, accessibility, accuracy, completeness, consistency, and currency, according to Laranjeiro et al. (2015). There are a number of well-known and important contributions that have been made with respect to the assessment of data quality and the corresponding dimensions in structured data (Batini et al., 2011; English, 1999; Lee et al., 2002; Pipino et al., 2002, Redman, 1997; Umbrich et al., 2015) and unstructured data (Immonen et al. 2015; Kiefer 2016; Kiefer 2019; Sonntag 2004). In the remainder of this section, the dimensions are shortly presented and discussed regarding their importance and influence on the quality in the context of Wikis. Furthermore, the currency metrics are discussed concerning their useability in Wikis and especially in Wikipedia since the main focus is on the quality dimension currency as the in chapter 5 proposed metric measures the currency.

3.2.1 Accessibility

While there are various definitions of accessibility, they all share the aim to measure whether the needed data is obtainable for users. (Bovee et al. 2003; Gardyn 1997; Caballero et al. 2014). It can be seen as the foundation of all other data quality dimensions (Bovee et al. 2003) since further measurements cannot be performed without retrieving the data in the first place. Some authors like Batini and Scannapieco (2016) and Huang et al. (2012) define accessibility in a broader context. Batini and Scannapieco (2016) view accessibility as the ability to access the data regardless of one’s culture, physical status, and available technology. To increase accessibility, they propose the use of simple language as well as short and fitting image descriptions. Other authors such as Huang et al. (2012) expand the definition of accessibility to the believability of data and providing an appropriate amount of information to its users. Nevertheless, there are only very few metrics that actually measure accessibility. Additionally, these metrics are only useful in a limited context such as Batini and Scannapieco (2016), who propose an accessibility metric to measure the percentage of words in a text that can be understood by an average person. While accessibility may appear as a fundamental dimension, in the context of Wikis specifically granting access to data can be seen as the main task of a Wiki (McAfee 2006; Seibert et al. 2011). For that reason, this paper will not discuss the accessibility of Wikipedia further. Its database contains over 50 million articles[3] and is available in 319 different languages[4], turning every individual owning an internet connection into a potential user.

3.2.2 Accuracy

Many authors define accuracy as the closeness of an attribute value stored in an IS to its real-world counterpart (Batini and Scannapieco 2016; Loshin 2011; Moraga et al. 2009; Naumann and Rolker 2000; Redman 1997). Batini and Scannapieco (2016) distinguish between two types of accuracy. The first one is called syntactic accuracy and focuses on the syntactical correctness of the data. Thus, usually, a poor syntactic accuracy occurs if there are spelling errors. For that reason, automatic spelling programs can be used as a tool to avoid syntactic accuracy and to improve simultaneously syntactical accuracy. Especially in

[3] https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

[4] https://en.wikipedia.org/wiki/List_of_Wikipedias

structured data, the values usually can only take values from a finite set $\mathcal{D}$. In such cases, it is possible to correct syntactic errors with distant functions, which measure how similar a value v having a syntactic error is to all allowed values in $\mathcal{D}$. For instance, how many characters differ in two words. The second type is called semantic accuracy and measures the closeness between the value in the IS and the real world. Semantic accuracy is mostly measured with 0-1-functions (also called indicator functions), that take the value zero if the data is semantically incorrect and one if the data is semantically correct. Even though, measuring accuracy is a complex and usually cumbersome process as it requires an evaluation of the real-world data value, it is an crucial data quality dimension in the context of Wikis as the trust of its users is a very important criterion for the success, which only be achieved by accurate data (Bhatti et al. 2018). However, there are no accuracy metrics for Wikis in the literature, leading to a system where its users need to have confidence in an editor's ability to conduct his research thoroughly to write accurate articles.

3.2.3 Completeness

Completeness describes whether all required information is present in an IS (Caballero et al. 2014; Loshin 2011; Naumann and Rolker 2000; Redman 2001; Wang et al. 1995a). In structured data, completeness can be assessed relatively simply. Either all required data is available in the IS or it is clear which data is missing. Therefore, completeness in structured data can be measured with 0-1-functions or as a percentage of necessary data available. Determining the reason why data is missing, however, is a more complex task. It can be caused by three different reasons. First, the data does exist but is unknown. Second, the data does not exist, or third, both combined, if the existence of the data is uncertain or cannot be confirmed (Firmani et al. 2016). The task of determining completeness in unstructured data is far more complex and difficile in some cases even practically impossible. Nevertheless, some authors such as Kiefer (2016) define it as an important data quality dimension in unstructured data. In literature, there are only a few existing approaches such as Arolfo and Vaisman (2018), who define an entity of unstructured data as complete when all metadata is present (i.e. a tweet and the corresponding author). However, this definition is very similar to the definition in structured data. Another approach was proposed by Cheng et al. (2010), who used natural language processing (NLP) to classify whether unstructured radiology reports contained sufficient information for tumour status classification. A generalisation of this approach to other unstructured data, especially Wikis seems difficult since even though it contains mostly unstructured data its completeness is challenging to ensure and hard to define. Large Wikis as Wikipedia define that an article must summarize the topic comprehensively[5]. However,

[5] https://en.wikipedia.org/wiki/Wikipedia:What_is_an_article%3F

Wikipedia's definition of comprehensiveness is not transparent and as a defining framework for articles probably insufficient to account for in all articles. Besides completeness of single Wiki articles, assessing the completeness of a whole Wiki, ensuring coverage of all important topics, is as essential. For instance, Wikipedia defines criteria that topics must fulfil to be covered in an own article such as significance and the availability of reliable sources[6]. Nevertheless, it is hardly possible to measure how complete a Wiki is since a determination for all possible topics whether they classify as relevant for the corresponding Wiki or not would be required.

3.2.4 Consistency

Consistency is defined as the degree of information being compatible with other similar information objects (Laranjeiro et al. 2015), nevertheless, there are different concepts of what exactly compatible means. Some authors such as Liaw et al. (2013) define consistency as the extent to which data is in a uniform format and easy to apply. Dufty et al. (2014) define consistency as the extent to which a data set complies with standard definitions. In comparison to other approaches, their definition has also a time-related component. The most widely used definition is that consistency measures the degree to which a data set fulfils predefined constraints (Benkhaled and Berrabah 2019 Cykana et al. 1996; Kumar and Thareja 2013). Batini and Scannapieco (2016) distinguish between two types of constraints. The first ones are called intrarelation constraints and are defined as constraints that regard single or multiple attributes in a relation. For instance, that age is a number between 0 and 120 or that a person aged below 15 has not graduated from university. Interrelation constraints are used in the context of multiple relations such as two different data sets one containing a curriculum vitae of a person with a bachelor's degree from 2018 and one list of graduates from her university in 2018. If the list of graduates contains the name of the said person, it would be consistent. In context of semi-structured and unstructured data only a few works focus on consistency. Blake and Mangiameli (2009) propose a metric based on the Porter stemming algorithm to measure consistency in customer data sets. Eppler and Muenzenmayer (2002) measured consistency on homepages of companies as percentage of pages on the website that are in correct style. However, in context of Wikis this definition only provides little insights since all structures are unitarily standardized for all sites in a Wiki. Instead, a metric, measuring whether information of different articles is free of contradictions, is of greater interest. But building such a metric probably requires advanced NLP methods. Besides delimiting consistency of accuracy in such cases is ambitiously challenging.

[6] https://en.wikipedia.org/wiki/Wikipedia:Notability

3.2.5 Currency

Regarding the measurement of currency as a quality dimension in structured data, one of the first and most renowned contributions was provided by Ballou et al. (1998). They define currency (referred to as timeliness by the authors) as a function of the age, the delivery time, the input time, and the maximum given shelf life of an attribute. Similar definitions are provided by Even et al. (2010) and Li et al. (2012). However, defining currency of Wikipedia articles, using the above-stated definition will meet with difficulties. First, attributes are not defined in unstructured data and thus it is not instructed how to apply these approaches. Second, there is no predefined shelf life for an article. Estimating the shelf life for every article would require a cumbersome manual process and therefore is expensive and time-consuming. Finally, real-world events reducing the shelf life would not be considered, as the shelf life is modelled as a constant per article. Another approach from Heinrich et al. (2007), Heinrich and Klier (2011), Heinrich and Klier (2015) as well as Wechsler and Even (2012) is the development of probability-based metrics. The metric in Heinrich et al. (2007) and Heinrich and Klier (2011) is based on the assumption of an exponential distribution and similar to Wechsler and Even (2012) who propose a metric based on Markov-Chains, which they assume to be memoryless and approximately exponential distributed. Heinrich and Klier (2015) propose a metric based on conditional expectation and additional metadata. On the one hand, the idea of a probability-based metric is valuable and transferable to the context of Wikipedia and thus serves as a starting point to define a metric for the currency of Wikipedia articles. On the other hand, the approaches focus on structured data and do not account for events that affect the currency of articles.

In the context of unstructured data, one of the most renowned contributions has been made by Batini and Scannapieco (2016). They define different data quality dimensions for different kinds of unstructured data as maps, texts, and pictures. In general, they state that currency is measured with respect to the latest time data it has been updated. A similar idea was proposed by Firmani et al. (2016). Zhu and Gauch (2000) also define currency (in the context of websites) as the most recent time an update was carried out on a homepage. However, these approaches do not account for the specifics of articles (e.g., the occurrence of events affecting the currency). For knowledge bases, Hao et al. (2020) implemented a classifier to decide if a fact in a knowledge base is outdated based on historical update frequency and the time of the existence of the fact. In summary important contributions have been made with respect to the currency of unstructured text data. However, they only take simple features like the age and average update frequencies as an approximation for currency and do not focus on real-world events. This can cause inaccuracies when using these currency metrics on Wikipedia articles since not all articles outdate in the same way. To give an example, the article about the binomial formula is most probably still up-to-date, even though its most recent update was performed in January 2020. However, an article dealing with the Covid-19 pandemic from the same time would be completely outdated a year later. While partly, this drawback can be compensated by additionally accounting for average update frequencies, non-regular events (e.g., the election of a person to be president) affect the updates needed in a possibly fundamental way and thus should be included.

3.2.6 Data quality metrics and their requirements

Data quality metrics provide measurements for data with greater metric values representing a greater level of data quality and each data quality level being represented by a unique metric value. According to Heinrich et al. (2018a), there are two main reasons data quality metrics are needed. First, to support data-based decision making under uncertainty. Especially, to indicate to which degree decision makers can rely on the underlying data. Second, the data quality metrics can be used to foster an economically oriented management of data quality.

Number	Requirement for data quality metrics
(R1)	Existence of minimum and maximum metric value
(R2)	Interval-scaled metric values
(R3)	Quality of the configuration parameters and the determination of the metric value
(R4)	Sound aggregation of the metric values
(R5)	Economic efficiency of the metric

Table 3: Requirements for data quality metrics (based on Heinrich et al. (2018a))

In the remainder of this subsection five requirements for data quality metrics based on the work of Heinrich et al. (2018a), which can be seen in Table 3, are presented. The first requirement (R1) is the existence of a minimum and maximum metric value. This includes that this minimum and maximum value can be attained. Thereby the maximum value represents a perfectly good data quality meaning full accordance with the data value stored in an IS and the real-world. On the other hand, the minimum value represents the highest level of imbalance between the real-world and the data stored in an IS that can be reached stating the margin cannot be further increased. Both these states are unique and therefore described by the minimum and maximum metric value. If (R1) is violated it may occur that the data quality is already unimprovable but due to the missing maximum value it is still tried to improve the quality. According to the second requirement (R2), the metric values need to be interval-scaled, in particular indicating that differences between two metric values can be determined and are meaningful. This is essential for data driven decision making if there are several different alternatives to choose from. If this requirement is disobeyed for example by using an ordinal scaled metric than there is no possibility to specify and compare the difference between two categories. This inability exacerbates the process of an efficient and beneficial improvement of the data quality. The next requirement (R3) focuses on the quality of the configuration parameters and the determination of the metric value. This means the configuration parameters of the data quality metric as well as the determination of the metric values is determined according to the quality criteria objectivity, reliability, and validity. Objectivity denotes the degree to which both the configuration parameters and the data quality metric values are independent of external influences such as interviews. The standards of objectivity are not met if the metric lacks a precise specification of procedures for determining the parameters and values. Reliability focuses on the replicability of the methods used to determine the parameters and the metric values. Reliability can be ensured by using statistical methods and correct database queries. In this case the results of the metric remain the same if multiple times applied to the same data set. Validity is defined as the degree to which a metric measure what it aims to measure. Thus, validity is violated if the determination of the parameters or the metric values contradicts the goal of the metric. This can be prevented by using consistent definitions and well-founded statistical estimations. Metrics that do not fulfil (R3) can lead to wrong decisions and serious problems when evaluating data quality improvement measures. This occurs due to the fact that if objectivity or reliability are violated, two applications of the metric can result in different metric values. If validity is violated the data quality metric values may not reflect the true state of the data quality. The fourth requirement (R4) is that the metric values are applicable to single data values as well as to sets of data values. Moreover, it is necessary that the aggregation of the data values lead to consistent metric values. Violations of this requirement can lead to different results especially if the size of the data set varies over time. The last requirement (R5) focuses on the cost of the application of the metric. It is important that the additional benefit of applying the metric outweighs the cost of determining the configuration parameters and the metric values. This appears especially critical if the configuration parameters are not directly available but need to be assessed before applying the metric.

3.3 Evaluating probability-based metrics

A probability-based metric fulfils the first two requirements (R1) and (R2) presented in the last subsection above. Moreover, as it is objective it partly fulfils (R3) as well. Therefore, a probability-based metric has many advantages compared with other data quality metrics. Thus, the in chapter 5 presented currency metric will be a probability-based metric as well. For this reason, in this section different methods for evaluation methods of probability-based metrics are addressed. There are two important categories to evaluate a probability-based metric. The first one analyses the reliability of the estimated probabilities. The second one focuses on the performance of the metric as a classifier and how to optimize it. Therefore, it is assumed to have two data sets. The first one $Y = \{Y_1, \dots, Y_n\}$ is called the verification. Y_i can only take two values, either zero (negative) or one (positive). In our proposed metric Y_i will be equal to one if a notable event has happened and zero if there was no notable event. However, for other applications, there are many different possible interpretations such as the state of a machine (broken or working) or if it is raining on a certain day or not. The second data set $X = \{X_1, \dots, X_n\}$ contains the probability forecasts. Thereby the forecast value X_i corresponds to the verification value Y_i. In comparison to Y_i, the possible range of values of X_i are all real numbers between zero and one. The forecast value X_i represents the probability that Y_i is equal to one.

3.3.1 Effectiveness of a probability-based metric with respect to the estimated probabilities

To be a useful metric, it must be ensured that the estimated probabilities correspond to the actually observed relative frequencies, which can be assessed in terms of reliability (Hoerl and Fallin 1974; Murphy and Winkler 1977; Murphy and Winkler 1987; Sanders 1963). A probability forecast is called reliable if the event actually occurs with an observed relative frequency consistent with the forecast value. This suggests, more specifically, that if we only consider instances i with $X_i = x$ ($x \in [0{,}1]$) the event happens with a relative frequency of x. Clearly, this definition is not very useful. Theoretically, there are uncountable many values x can take. Therefore, the probability of two different instances X_{i_1}, X_{i_2} having the same probability is equal to zero. Even accounting for numerical limits in the amount of numbers a computer can represent it is still improbable to observe the same value x twice or even more frequent. But if a certain number x is only observed once in the data set X the relative frequency is not meaningful. To avoid this problem reliability plots are used.

Reliability plots

The following description is based on the work of Bröcker and Smith (2007). In a reliability plot, the observed relative frequencies are plotted against the forecast values. To evade the problems described above, the forecast values are divided into K bins $B_1, \dots, B_K$. Each bin represents a subinterval of the interval $[0{,}1]$. Moreover, it must hold that the bins are mutually exclusive ($B_i \cap B_j = \emptyset \; \forall i \neq j$) and collectively exhaustive ($\bigcup_{j=1}^{K} B_j = [0{,}1]$). A natural and frequently used method is to choose the bins equidistantly. However, it is not set as a requirement, especially for non-uniform distributed forecast values it is possible to define the bins as equally populated. The selection of K is as the definition of the bins not strictly defined. However, it must be ensured that all bins contain enough samples. All forecast values X_i are sorted into the corresponding bin B_k (such that $X_i \in B_k$). Therefore, $I_k = \{i \in \{1, \dots, n\}: X_i \in B_k\}$ is defined as the collection of all indices i, such that X_i falls in the interval B_k ($k \in \{1, \dots, K\}$).

Furthermore, the observed relative frequency f_k of the bin B_k is defined as

$$f_k = \frac{\sum_{i \in I_k} Y_i}{|I_k|},$$

with $|I_k|$ denoting the numbers of elements in the set I_k. Moreover, it is necessary to calculate a typical forecast probability for each bin B_k. A very elementary approach would be to choose the arithmetic centre of the bin B_k. However, an evident disadvantage of this selection is that the arithmetic centre not necessarily coincidences with the average value of the forecast values in B_k. Especially if the values in a bin B_k are heavily skewed towards one side, this could lead to wrong conclusions about whether the forecast probabilities are reliable. Thus, a better choice for a typical forecast probability of the bin B_k is

$$r_k = \frac{\sum_{i \in I_k} X_i}{|I_k|},$$

the average forecast value of the bin B_k. In a reliability plot f_k is plotted against r_k for all bins B_k. Figure 3 shows a typical reliability plot, where three different probability forecasts are compared. A perfectly reliable probability forecast would follow the diagonal line $y = x$.

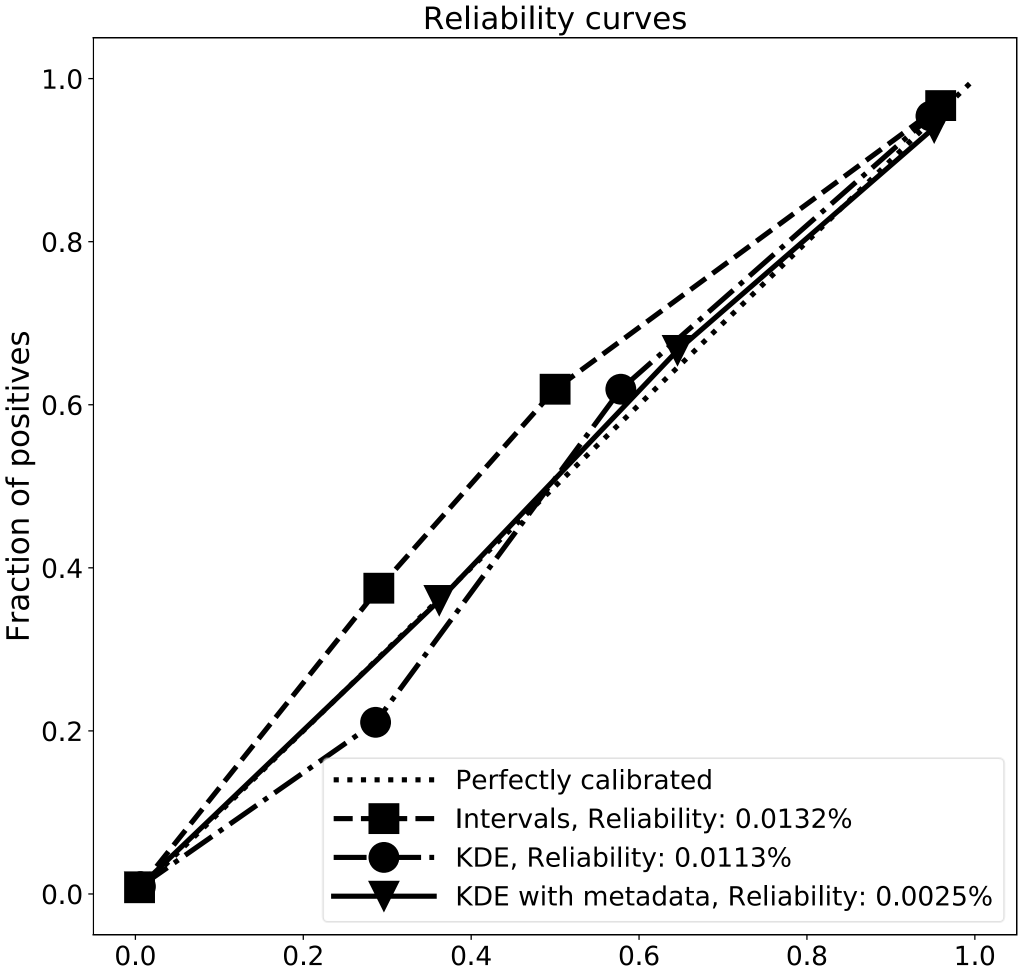

Figure 3: Example of a reliability curve with three approaches (Heinrich et al. 2018)

Obviously, this scenario is almost unreachable in practical applications. Thus, a probability forecast is called reliable if the corresponding probability curve follows the diagonal line closely. If there are several probability forecast methods, Murphy (1973) proposed a reliability score R, with

$$R = \frac{1}{n}\sum_{k=1}^{K} |I_k|(f_k - r_k)^2,$$

being the mean squared deviation from the diagonal weighted by the number of test cases in each bin. Therefore, the smaller the value of the reliability score the smaller the discrepancy between the estimated probabilities and the actually observed relative frequencies.

Receiver operating characteristic curve

The receiver operating characteristic (ROC) curve was first used during World War II in the analysis of radar signals. For the purpose of increasing the prediction of correctly detected Japanese aircraft from radar signals after the attack on Pearl Harbor, the United States army measured the ability of radar receiver operators to detect bombers of the enemy (Collinson 1998) using the receiver operating characteristic.

The following introduction is based on the work of Fawcett (2006). Firstly, the absence of a clear dissociation whether a ROC curve evaluates the reliability of the estimated probabilities or the classification needs to be taken into account as it can be used for both. Nevertheless, it is presented in this subsection to be consistent with other publications such as Heinrich et al. (2018b). To obtain the ROC curve for each probability value $x \in X_0 = X \cup \{0\} = \{0, X_1, \dots, X_n\}$ the set $Y = \{Y_1, \dots, Y_n\}$ is divided into two subsets. In this case x is called the cut-off point. The set $I_p^x = \{i \in \{1, \dots, n\} : X_i > x\}$ contains all indices i, such that Y_i would be classified as one (positive) by a binary classifier with threshold x. Conversely, the set $I_z^x = \{1, \dots, n\} \setminus I_p^x = \{i \in \{1, \dots, n\}: X_i \leq x\}$ contains all indices i such that Y_i would be classified by the same classifier as zero (negative). Next the true positive rate *tpr* (also called recall or sensitivity and discussed in subsection 3.3.2 in more detail) is calculated using the following formula

$$tpr^x = \frac{\sum_{i \in I_p^x} Y_i}{\sum_{i=1}^{n} Y_i}.$$

As the name states the true positive rate compares the number of positives that were classified as positives in comparison to the number of all positives. Moreover, the false-positive rate fpr needs to be calculated with the formula

$$fpr^x = \frac{|I_p^x| - \sum_{i \in I_p^x} Y_i}{|Y| - \sum_{i=1}^{n} Y_i},$$

which is the rate of negatives that are incorrectly classified in comparison to all negatives. Instead of using the term false positive rate it sometimes is described as $1 -$ specificity^x. In such cases, specificity is defined as $\frac{|I_z^x|-\sum_{i\in I_z^x} Y_i}{|I_z^x|}$. It can be shown that it is actually $1-fpr^x$. The ROC curve can be obtained by plotting tpr^x against fpr^x for all $x \in X_0$ resulting in a step function. Figure 4 shows an example of a ROC curve. The diagonal line $x = y$ represents a classifier guessing the class randomly. For instance, if the classifier randomly guesses the positive class and the negative class with the same probability, it can be expected that it classifies half of the positives correct and half of the negatives incorrect. Thus, it yields the point $(0.5, 0.5)$. If these numbers are varied such as that the classifier guesses the positive class only 10% of the time, it is expected that only 10% of the positives are correctly classified. On the other side, only 10% of the negatives are expected to be false classified yielding in the point $(0.1,\ 0.1)$. In comparison, a perfect probability-based classifier would result in a straight line from the origin $(0, 0)$ to the upper left corner $(0,\ 1)$ and from that point, it would draw a line to the opposite right corner $(1,\ 1)$. In such a case using the optimal probability value x_o as the cut-off point (the value, resulting in the point $(0,\ 1)$), would lead to a perfect classifier, identifying all values correctly (as all positives are classified as positive and no negative as positive).

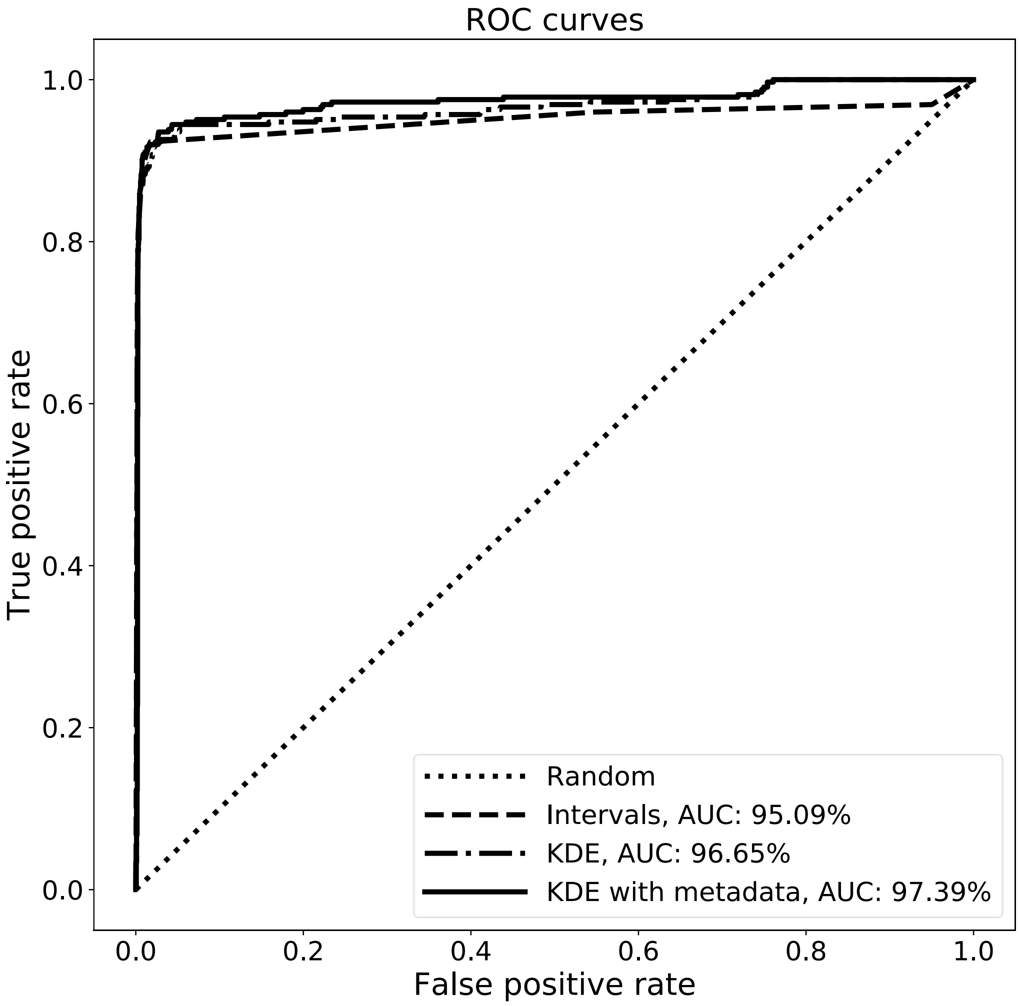

Figure 4: Example of a ROC curve for three different approaches (Heinrich et al. 2018)

For cut-off point $x > x_o$, still, all of the negatives are classified correctly (as all of the negatives have a corresponding probability $X_i < x_o$ to be a positive, but some of the positives are classified incorrectly as negatives. For $x = \max_{i \in \{1, \ldots, n\}} X_i$, the classifier always guesses negative. Thus, all negatives are correctly classified, but no positive is classified correctly yielding in the point $(0, 0)$. This explains the straight line from $(0, 0)$ to $(0, 1)$. Similarly, the straight line from $(0, 1)$ to $(1, 1)$ of a perfect classifier can be explained by decreasing the cut-off point to zero. In general, the ROC curve always starts in $(0, 0)$ and ends in $(1, 1)$. Due to the premise that the ROC curve of a probability-based classifier is supposed to be above the angle bisector ($y = x$) in cases, this prerequisite is not fulfiled a random classifier is a superior choice. However, only having the ROC curve it is sometimes difficult to decide which probability-based approach is superior. For a deeper understanding of comparing approaches in the next paragraph, a metric will be defined.

Area under the curve

As above this concept is explained following the work of Fawcett (2006). The area under the curve (AUC) is as the name states the area under a ROC curve. Taking values between zero and one, a value of one is seen as a perfect classifier (as the unit square has an area of one). The guessing classifier would achieve an AUC value of 0.5. Thus, a reasonable classifier should always have an AUC value greater than 0.5. The AUC has a probabilistic interpretation as it is the probability that the probability-based metric assigns a higher probability of being a positive to a randomly chosen $Y_j \in \{Y_i \in Y : Y_i = 1\}$ than to a randomly chosen $Y_k \in \{Y_i \in Y : Y_i = 0\}$. A proof for this can be found in Hand (2009). Since the ROC curve is a step function the AUC can be calculated effortlessly by just iteratively adding trapezoids[7]. The ability to compare different probability-based metrics is a huge advantage of the AUC. Hosmer et al. (2013) have presented a scale to assess the quality of probability-based metrics. They describe an AUC greater than 0.7 as acceptable discrimination. If the AUC value is greater than 0.8 the discrimination is called excellent, while an AUC value greater than 0.9 is named outstanding discrimination. However, some authors criticize the AUC value as easy to manipulate and ignoring the goodness-of-fit model (Lobo et al. 2008). Nevertheless, the AUC is a widely accepted metric to assess the quality of probability-based metrics (Hand 2009).

3.3.2 Effectiveness of a probability-based metric with respect to classification

As described in the section above the probability-based metric also can be used as a classifier. For this, a certain cut-off point $x \in [0,1]$ is determined and then all Y_i with $X_i > x$ are classified

[7] Trapezoids are used instead of rectangles to account for the uncertainty resulting from the fact that only finitely many instances can be observed, and the actual ROC curve would be a smooth function, although rectangles would be more precise for a single ROC curve.

as positives and vice versa all Y_i with $X_i \leq x$ are classified as negatives. An obvious choice for x would be 0.5. However, a method to determine an optimal cut-off value is presented further down in this subsection. Firstly, for a better understanding, classical metrics for assessing the quality of (binary) classification are described. In the following the set $Y_r^p = \{Y_i \in Y : Y_i = 1\}$ is called the real-world positives, $Y_r^n = \{Y_i \in Y : Y_i = 0\}$, the real-world negatives, $Y_c^p = \{Y_i \in Y : X_i > x\}$ the classified positives and $Y_c^n = \{Y_i \in Y : X_i \leq x\}$ the classified negatives.

Evaluation of binary classifiers

The following paragraph is based on the work of Powers (2011). The start of each evaluation of a binary classifier is a confusion matrix occasionally called contingency table, which is a 2×2 matrix.

		True condition	
		Condition positive	Condition negative
Estimated Condition	Predicted condition positive	True Positive (TP)	False positive (FP) (Type I error)
	Predicted condition negative	False negative (FN) (Type II error)	True Negative (TN)

Table 4: Structure of a confusion matrix

Table 4 shows the structure of a confusion matrix. True positives (TP) describe the number of instances that are positives in the real-world and are classified as positives $\left(\left|Y_c^p \cap Y_r^p\right|\right)$. Similarly, True Negative (TN) is defined as the number of instances that are negative in the real-world and classified as negatives $(|Y_c^n \cap Y_r^n|)$. Together they represent all correctly classified instances. Conversely, the number of instances that are wrongly classified as positives as they are negatives in the real-world $\left(\left|Y_c^p \cap Y_r^n\right|\right)$ are called False Positives (FP). Furthermore, False Negatives (FN), describes the instances that are positive in the real-world and are falsely classified as negatives $\left(\left|Y_c^n \cap Y_r^p\right|\right)$. To determine and assess the level of quality of a classifier **accuracy** is one of the most promising. It is defined as the rate of correctly classified instances to the total number of instances and therefore can be calculated by

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}.$$

However, accuracy does not yield any interpretation of marginal properties. For that reason, it is frequently seen as not very meaningful. Especially if the real-world data is heavily skewed (for instance 95% of all Y_i are negatives) accuracy can be a misleading metric, as in such a case a bad classifier that always guesses the skewed category can achieve high accuracy (in the example above a classifier that always predicts negative gets an accuracy of 95%). Thus, in practice, there is also a need to assess the quality of a classifier with respect to the marginal predictions. In this context, three metrics are especially important, **recall, precision,** and the **F₁** measure. These three metrics are computed for positives and negatives each. In other scientific fields such as medicine recall sometimes is referred to as sensitivity. Moreover, in the literature the terms recall, and precision are sometimes only used in the context of the positive class, while the recall of the negative class is called true negative rate or specificity and the precision of the negative class is called negative predictive value. However, in this paper, the recall of the class a (with $a \in \{positive, negative\}$) is defined as

$$recall_a = \frac{|Y_r^a \cap Y_p^a|}{|Y_r^a|}.$$

A probability-based classification model with a high recall of class a is able to classify most of the instances of the real-world class a as belonging to this class. However, the recall is easy to manipulate as a classifier, which always classifies as class a will achieve a recall of one, which is the highest and best value for a recall. Thus, to balance this the precision of class a is defined as

$$precision_a = \frac{|Y_r^a \cap Y_p^a|}{|Y_p^a|}.$$

High precision of the class a shows that most of the as a classified instances are belonging to the class a. Often recall and precision are opposing metrics, meaning increasing one leads to a decrease of the other. Depending on the classification one of these metrics can be more important to optimize. If the goal is to minimize false positives a high precision (of the class p) is necessary, for instance, a spam-filter (as it should not misidentify legitimate e-mails as spam). On the other hand, if it is necessary to minimize false negatives a high recall (of the class p) must be achieved. A possible scenario would be a cancer screening. In practice, however, there are many cases in which both metrics are of interest. In such a case the F_1^a measure of class a defined as the harmonic mean of $recall^a$ and $precision^a$

$$F_1^a = 2\frac{recall^a \times precision^a}{recall^a + precision^a}$$

is of particular interest. A high F_1 measure of the class a means that the recall, as well as the precision, are high. This is the optimal case as it shows that the model classifies most of the instances that belong to the class a correctly, but also does not classify many instances that not belonging to the class a as a.

Youden's J statistic to find the optimal cut-off point

Youden's J statistics for a binary probability-based classifier is defined as

$$J = recall^p + recall^n - 1.$$

It was first presented by the Australian-born American statistician William J. Youden in 1950 to capture the performance of a classifier in a single statistic (Youden 1950). This presentation of Youden’s J statistic is based on the work of Youden (1950) and Schisterman et al. (2005). Youden’s J statistic is defined such that $J \in [-1, 1]$, with $J = -1$ representing a classifier that classifies every instance incorrect and $J = 1$ representing a perfect classifier. A classifier that classifies randomly is expected to achieve $J = 0$, as in this case there are the same proportion of positives are classified as positives as are classified as negatives. In relation to a probability-based metric Youden’s J-statistic is used to find the optimal cut-off point for classification. For this, J_x is computed for all $x \in X_0$ and the optimal cut-off point is defined as

$$\hat{x} = \underset{x \in X_0}{\operatorname{argmax}} J_x.$$

Using the cut-off point $\hat{x}$ resulting in the best possible combination of classifying a high percentage of positives as positives, while simultaneously classifying a high percentage of negatives as negatives. Hence, resulting in a favourable classifier compared to a classifier using other cut-off values $x \in X_0$. J_x can be interpreted graphically in the ROC curve as the height of the ROC curve above the diagonal line $y = x$. As a result, the probability-based classifier using the optimal cut-off point $\hat{x}$ can be interpreted as a classifier that is the farthest away from a randomly guessing classifier.

4 Related Work

As shown in the last chapter important contributions with respect to the assessment of different data quality dimensions and especially currency have been made. However many of these general approaches are derived for structured data and thus cannot directly be applied in the context of unstructured data (Batini and Scannapieco 2016; Kiefer 2019). Moreover, it was discussed that existing currency metrics in the context of unstructured data cannot be adapted for the use in Wikis. Therefore, the remainder of this chapter is offering an overview of existing approaches and metrics for the quality assessment of Wikis, especially Wikipedia (as quality metrics in Wikis are usually designed and evaluated through the example of Wikipedia) are presented and discussed. Figure 5 provides an overview of the different approaches used in this context. The most widely used idea is to classify the articles in different categories, which often refer to the Wikipedia intern grading system of the articles[8] or to classify the quality of single edits. Due to the popularity and prevalence, Di Sciascio et al. (2017, 2019) developed a web tool called Wikilyzer, which allows everybody to create their own classification rules for Wikipedia articles. Subsequently, these attempts are judged and compared to a number of state of the art methods for classifying Wikipedia articles, which are presented in the following.

[8] https://en.wikipedia.org/wiki/Wikipedia:Content_assessment#Grades

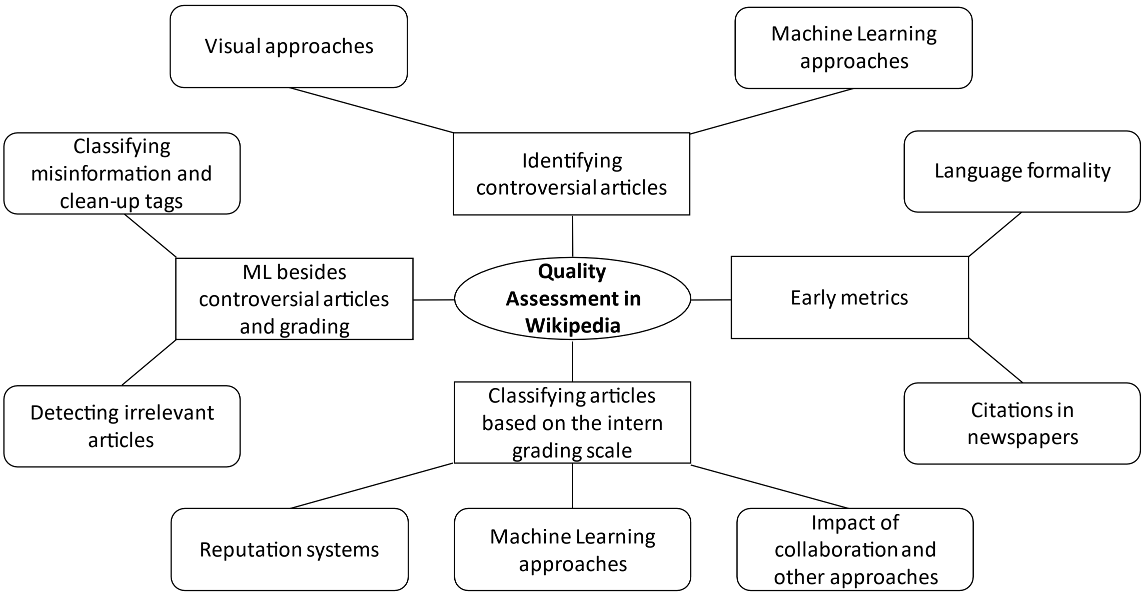
Visual approaches
Machine Learning approaches
Classifying misinformation and clean-up tags
Identifying controversial articles
Language formality
ML besides controversial articles and grading
Quality Assessment in Wikipedia
Early metrics
Detecting irrelevant articles
Classifying articles based on the intern grading scale
Citations in newspapers
Reputation systems
Machine Learning approaches
Impact of collaboration and other approaches

Figure 5: Approaches for quality assessment in Wikipedia

For a better understanding Table 5a contains all contributions in the context of classifying articles based on the internal grading scale that are presented in the following. A very popular approach in classifying the quality of Wikipedia articles is to implement reputation systems for authors. The basic idea of such a system is to assess the quality of an editor based on edits he previously made and thereafter evaluating articles based on the quality of its editors. One of the first and most renowned contributions was made by Zeng et al. (2006), who developed a reputation system based on a Bayesian network with a beta distribution to compute the quality of fragments of articles based on the trustworthiness of its editors to classify the articles in the classes featured articles and clean-up-articles. Adler and Alfaro (2007) implemented a content-driven reputation system, where the reputation of an editor A depends on the amount of his text and edits, which is preserved after an edit of another author B and the reputation of the author B. This system allowed them to grade the quality of single edits, especially with respect to the longevity of the edit (implying an edit was not reverted retroactively). Similar ideas can be found in Lim et al. (2006) and Qin and Cunningham (2012). The approach of Lim et al. (2006) is to calculate a trust value for each article and each editor, depending on one another. It is used to assess the quality of the articles based on a self-defined quality scale. However, assessing the quality of this approach can only be roughly estimated since the instantiation of this quality scale is not transparent. Qin and Cunningham (2012) compared an approach very similar to Adler and Alfaro (2007) with an editor authoritativeness metric based on the centrality in networks and metric that combines both approaches. All three metrics were used to classify articles based on the Wikipedia intern grading scale.

Main idea	Category	Contributions in this field
Classifying articles based on the intern grading scale	Reputation systems	• Adler and Alfaro (2007) • Javanmardi et al. (2010) • La Robertie et al. (2015) • Lim et al. (2006) • Qin and Cunningham (2012) • Suzuki and Yoshikawa (2013) • Wöhner et al. (2015) • Zeng et al. (2006)
	Machine Learning approaches	• Dalip et al. (2009) • Dalip et al. (2014) • Dalip et al. (2017) • Dang and Ignat (2016a) • Dang and Ignat (2016b) • Dang and Ignat (2016c) • Dang and Ignat (2017) • Halfaker (2017) • Shen et al. (2017) • Warncke-Wang et al. (2013) • Zhang et al. (2018) • Zhang et al. (2020)
	Impact of collaboration and other approaches	• Blumenstock (2008a) • Blumenstock (2008b) • Brandes et al. (2009) • Di Sciascio et al. (2017) • Di Sciascio et al. (2019) • Kittur and Kraut (2008) • Li et al. (2013) • Stvilia et al. (2005b) • Warncke-Wang et al. (2015) • Wilkinson and Huberman (2007) • Wöhner and Peters (2009)

Table 5a: Contributions in classifying articles based on the intern grading scale

Javanmardi et al. (2010) proposed a reputation system to classify editors into admins and vandals. Their approach is based as well on the longevity of the edits from an author. All of the approaches have in common that they are sensitive to “edit wars”, where two or more editors constantly delete changes of the other editor(s). In such cases, the reputation of all editors shrinks as their versions are constantly deleted. Such a system would punish editors, who are in an edit war with vandals, who try to damage some articles on purpose. Suzuki and Yoshikawa (2013) therefore present a complex reputation system that is resistant to edit wars. To achieve this the authors used non-linear functions to assess the quality of editors and edits. With this model, they classified articles into the categories featured and good articles. La Robertie et al. (2015) provided a reputation system based on the mutual reinforcement principle to discriminate between different quality levels of Wikipedia articles. However, Wöhner et al. (2015) proved that the quality of articles not necessarily depends on the reputation of its authors and that the quality of a reputation system is highly dependent on the definition of reputation.

Besides reputation systems, many authors try to classify articles with other approaches. One of the first contributions in this context was made by Stvilia et al. (2005a), who showed that featured articles can be distinguished from other articles by their significantly better-developed discussion pages. Similarly, Wilkinson and Huberman (2007) showed that featured articles can be discriminated from other articles by a high number of editors per article. This conclusion disputes the results of Kittur and Kraut (2008) and Warncke-Wang et al. (2015) who studied how collaboration influences the quality of articles. They both obtained that a multitude of editors tends to decrease the quality of an article. As an avoidance strategy editors need to coordinate. Brandes et al. (2009) developed a bipolarity index based on a weighted directed network, which models how much text editors add, delete and restore. This bipolarity index allows a distinction between featured articles and controversial articles. Wöhner and Peters (2009) built a classifier based on different lifecycle metrics, especially focusing on the number of edits in the last months. This classifier differentiates the articles into three classes: featured articles, good articles, and articles for deletion. A very similar approach with similar results was presented by Li et al. (2013) as well.

Due to the fast development in machine learning methods in recent years, this approach was also focused as an important research frontier in classifying Wikipedia articles. Dalip et al. (2009) used a support vector machine with around 60 input features in three categories (text features, review features, and network features) to assess the quality of articles based on the intern grading system of Wikipedia. After further development, Dalip et al. (2014) presented an advanced method, where they were able to get comparable results with fewer features by using a genetic algorithm for the feature selection. Warncke-Wang et al. (2013) used the quality indicators presented by Stvilia et al. (2005b) along with some self-defined indicators as input features for a decision tree and a random forest to predict the grades of Wikipedia articles. Dang and Ignat (2016c) compared a variety of different machine learning approaches like logistical regression, k nearest neighbours, classification and regression trees, support vector machines, and random forests to predict the grade of an article with a focus on readability indexes. A further approach from Dang and Ignat (2016b, 2016a) used a deep neural network with the Doc2Vec of each article as input to predict the intern grade of the article. The performance of this approach was increased by using a recurrent neural network and long-short term memory (LSTM) (Dang and Ignat 2017). A slight performance improvement of this approach was achieved by Shen et al. (2017) employing a Bi-LSTM instead of an LSTM. A critical weakness of this approach is its use of former version's features, rendering the grade assessment for a single article expensive and time-consuming. Therefore, Zhang et al. (2018) provide a model based on text data and metadata of the current version alone. To assess the grade of an article they use a recurrent neural network. Zhang et al. (2020) defined three categories on their own (stalled articles, plateaued articles, and sustained articles) based on the intern grading system of Wikipedia and used three classes of variables (article attributes, article editing activities, and editors' attributes). Through time-series clustering methods and logistic regression, they were able to show that the quality of an article depends on attributes inherent to the article like topic importance, as well as on the editors of the article. Using the approach of Warncke-Wang et al. (2013) with slight modifications Wikipedia itself implemented the machine learning tool ORES[9] to score the quality of single edits and entire articles (Halfaker 2017). However, ORES is mainly used to assess the quality of new articles and the coverage of long-time existent articles is inconsistent.

A common problem of all classification methods is that the ground-truth quality grades of the articles are determined by the users/authors of Wikipedia and therefore are subject to subjective bias and missing comparability. Jemielniak and Wilamowski (2017) examined featured and good articles from different language version of Wikipedia and were able to show that there are huge differences in these criteria depending on an articles language. Moreover, the grades are past-oriented. However, due to the fast-paced world articles which used to be of high quality are possibly outdated now. Furthermore, featured articles can be easily distinguished from other articles as Blumenstock (2008a, 2008b) showed, who managed to get a precision of more than 96% just by classifying articles with more than 2000 words as featured. On the one hand, his findings are expected since featured articles gain close attention and are regularly extended and updated by various authors. On the other hand, the high accuracy of a classification using a simple metric like the article length questions the validity

[9] https://www.mediawiki.org/wiki/ORES

and usefulness of the label “featured article” in terms of a quality indicator. Moreover, less than 0.1% of the articles are featured[10], additionally hampering the use of this label as a quality metric, as the remaining 99.9% also need to be assessed regarding their quality.

Table 5b gives an overview of further contributions of assessing the quality of Wikipedia. These approaches focus on the number of edits, editors, and the number of citations in online newspapers of an article (Lih 2004) as well as the formality of the language in articles (Emigh and Herring 2005). However, these metrics are from the beginning of Wikipedia and probably cannot be applied in the same way today. A class of articles prone to bad quality are controversial articles. Therefore, some contributions focus on classifying such articles.

Main idea	Category	Contributions in this field
Early metrics	Language formality	• Emigh and Herring (2005)
	Citations in newspapers	• Lih (2004)
Identifying controversial articles	Visual approaches	• Viégas et al. (2004)
	Machine Learning approaches	• Bykau et al. (2015) • Jhandir et al. (2017) • Zielinski et al. (2018)
Machine Learning approaches besides controversial articles and grading	Classifying misinformation and clean-up tags	• Anderka and Stein (2012a) • Anderka et al. (2012b) • Sinanc and Yavanoglu (2013)
	Detecting irrelevant articles	• Ofek and Rokach (2015)

Table 5b: Contributions besides classifying articles based on the intern grading scale

One of the first contributions in this context was made by Viégas et al. (2004), who developed a visual tool displaying the development of an article over time, based on which sentence was written from which editor. Due to this tool, controversial articles could quickly be identified by an individual looking at the article visualization. An automatic approach was presented by Bykau et al. (2015) and Jhandir et al. (2017), who classified articles in controversial and non-controversial based on the edit histories. Zielinski et al. (2018) provided a machine learning approach based on the now discontinued article feedback tool to assess the controversy of articles, which was also able to determine controversial categories.

[10] https://en.wikipedia.org/wiki/Wikipedia:Featured_articles

Nevertheless, it needs to be taken into account that overall there is only a limited number of controversial articles, where this metric can be applied to. Anderka et al. (2012a, 2012b) presented an approach where they used a support vector machine to classify if an article contains a clean-up tag. These tags are set manually by Wikipedia users if they find a quality flaw without being capable of improving it themselves. This includes tags like missing citations and empty sections. Sinanc and Yavanoglu (2013) propose a model to find edits that deliberately damage an article by adding misinformation. For that purpose, the authors use an artificial neural network. To automatically detect articles that can be deleted due to missing relevance, Ofek and Rokach (2015) build a model based on a logistic model tree and a Bayes network. Overall, most of the in this paragraph described approaches are either not applicable for the nowadays version of Wikipedia or include only on a small number of articles.

Remarkably few papers focus on the currency of articles, despite Wong (2016) describes it as one of seven quality dimensions of encyclopaedias and Mesgari et al. (2015) found that currency is the least researched quality dimension of Wikipedia. In literature, other data quality dimensions are examined at least twice as much, while reliability is studied more than four times as much as currency. Moreover, the studies that focus on currency are restricted to a very specific knowledge domain. The examined articles indeed are up-to-date. However, in another study, Hatcher-Gallop et al. (2009) showed that in the medical area there are articles, which are not up-to-date, even as Wikipedia has an advantage in terms of currency in comparison to other online collaboration projects due to the high amount of voluntary editors. Dalip et al. (2017) described currency (resp. timeliness) as a pragmatic quality dimension, which can be measured with indicators like edit history and text context. The goal of the authors was to implement a machine learning approach to classify the articles based on the internal grading system of Wikipedia, so the currency of an article is not an output factor of the model and thus does not constitute a metric measure currency in Wikipedia articles. To the best of our knowledge, only two contributions are proposing such a metric. Tran and Cao (2013) used a pattern-based fact extraction approach to find outdated facts in the information boxes (infoboxes) in the upper section of a Wikipedia article. To accomplish this, they searched through other websites to find more current data for the corresponding facts. This approach however is limited to the infoboxes where the data is structured in a predefined way. For complete articles (including their textual part), this approach cannot be applied. Stvilia et al. (2005b) measured currency as the point in time an article was last updated. Their findings included that featured articles are more likely up-to-date (on average 3 days since the last revision) than other randomly picked articles (on average 46 days since the last revision). Albeit this approach is a good starting point, it has two major weaknesses. First, a majority of revisions in Wikipedia are due to minor fixes like spelling errors. Thus, this definition of currency could be biased by such small changes in an article. Second, it measures currency linearly and does not account for possible events leading to the instant obsolescence of an article. This can be illustrated by comparing articles of individuals, who died a long time ago and therefore will be subject to very few updates, compared to living persons, who experience events (e.g., being elected as president), which lead to updates required in their articles.

To conclude, existing approaches for currency metrics for both structured and unstructured data have only limited applicability on measuring the currency of Wikipedia articles. Even though many contributions are focusing on the quality of Wikipedia articles, only a small number of them focusses on currency metrics for Wikipedia. However, these approaches are either too simple to work on a wide range of articles as they only take the last update time into account or can only cope with small parts of an article like infoboxes. To address this research gap, we propose a probability-based event-driven metric for the currency of Wikipedia articles, that can account for real-world events, resulting in a need for an update.

5 A novel approach for measuring currency

As shown in the previous chapter there is a lack of currency metrics for Wiki articles that they can cope with the complete article and are not too simple to work on a wide range of articles, in this chapter a novel approach for measuring the currency of Wiki articles is proposed. The proposed metric avoids the issues and disadvantages of currently existing approaches. First, the basic idea is explained, before in the second section the mathematical foundation of the approach is illustrated. The last two subsections describe how to compute and estimate the probabilistic values that occur in the probabilistic formulas.

5.1 General idea

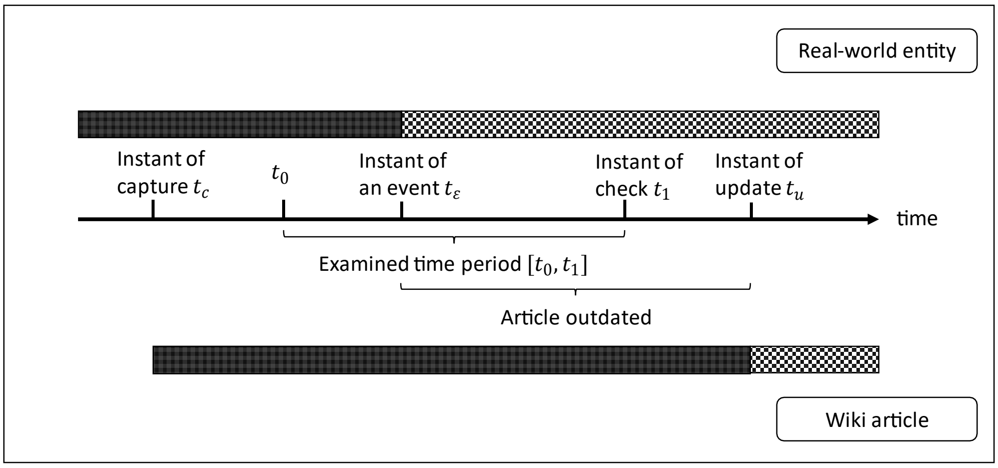

Figure 6: Basic concept

We consider a real-world entity e^r and a corresponding Wiki article e^d. Figure 6 shows the basic concept of modelling currency in Wiki articles. We assume that at the instant of capture t_c the Wiki article e^d perfectly reflects the state of the real-world entity e^r implying the article is up-to-date. The instant of capture can be interpreted as the time a Wiki article was created or last updated. At the instant t_ε an event ε occurs that changes the state of the real-world entity e^r. As a result, the Wiki article e^d does not further perfectly reflect the state of the real-world entity e^r. Thus, the article is outdated and needs to be edited to be up-to-date once again. Our approach aims to measure if there is an event ε that changes the state of the real-world entity e^r in the interval $[t_0, t_1]$. It is assumed that the Wiki article is up-to-date at the time point t_0, the first instant that is checked for events. In contrast to the few existing approaches measuring currency in Wiki articles, our approach has two major advantages. Firstly, the real-world entity e^r is modelled in a way that our approach can cope with the complete real-world entity e^r. Therefore, the approach is not limited to single parts of the Wiki article like infoboxes. Moreover, the currency is not measured linearly. As a consequence, this approach can cope with unregular events.

5.2 An Event-driven approach for measuring currency in Wiki articles

We define the currency of the real-world entity in the interval $[t_0, t_1]$ as

$$CUR(e_i^r, t_0, t_1) = \begin{matrix} false, \exists \varepsilon: t_\varepsilon \in [t_0, t_1] \\ true, \text{ else} \end{matrix}$$

It is important to state, that observing such events directly is not possible. Therefore, we need to define an indicator function for such events. In this paper, we define an indicator function, as a function that

1. is defined on a partition $\{u_1, \ldots, u_m\}$ of time instants in the time period $[t_0, t_1]$ and is mapped into the real numbers and
2. has an amplitude in the case of an event ε.

Examples of such indicator functions could include inter alia the daily pageview or the hourly number of search requests. According to the definition above we can define a notable event as an outlier of the indicator function since an outlier is defined as a point that significantly differs from the other observations (Grubbs 1969, Maddala and Lahiri 1992). We further define $O(I,\ e_i^r, t_0, t_1)$ as the set of all outliers in the interval $[t_0, t_1]$.Therefore in the case that our indicator is appropriate we have the following equivalence:

$$CUR(e_i^r, t_0, t_1) = false \Leftrightarrow \exists O(I, e_i^r, t_0, t_1) \neq \emptyset.$$

Defining $t_0 = s_0 < \cdots < s_l = t_1$ as arbitrary partition the following probabilistic formula holds:

$$P(CUR(e_i^r, t_0, t_1) = true) = P(O(I, e_i^r, t_0, t_1) = \emptyset) = 1 - P(O(I, e_i^r, t_0, t_1) \neq \emptyset)$$
$$= 1 - P(O(I, e_i^r, s_0, s_1) \neq \emptyset) \cdot P(O(I, e_i^r, s_1, s_2) \neq \emptyset | O(I, e_i^r, s_0, s_1) \neq \emptyset) \cdot$$
$$\ldots \cdot P\big(O(I, e_i^r, s_{l-1}, s_l) = \emptyset \big| O\big(I, e_i^r, s_{0,} s_{l-1}\big) = \emptyset\big).$$

This means the formula contains two kinds of probabilities, the first one $P(O(I, e_i^r, s_0, s_1) \neq \emptyset)$ is an unconditional probability and can be calculated relatively plain as we will show below. The second one, however, is a conditional probability and therefore needs more explanation.

A conditional probability is needed because it cannot be assumed that the sub-time periods are independent. In this case, the condition would be redundant as the conditional probability would equal the unconditional probability. However, the assumption of independence is always a very strong one and therefore needs clear and correct justification, which in our approach is not possible.

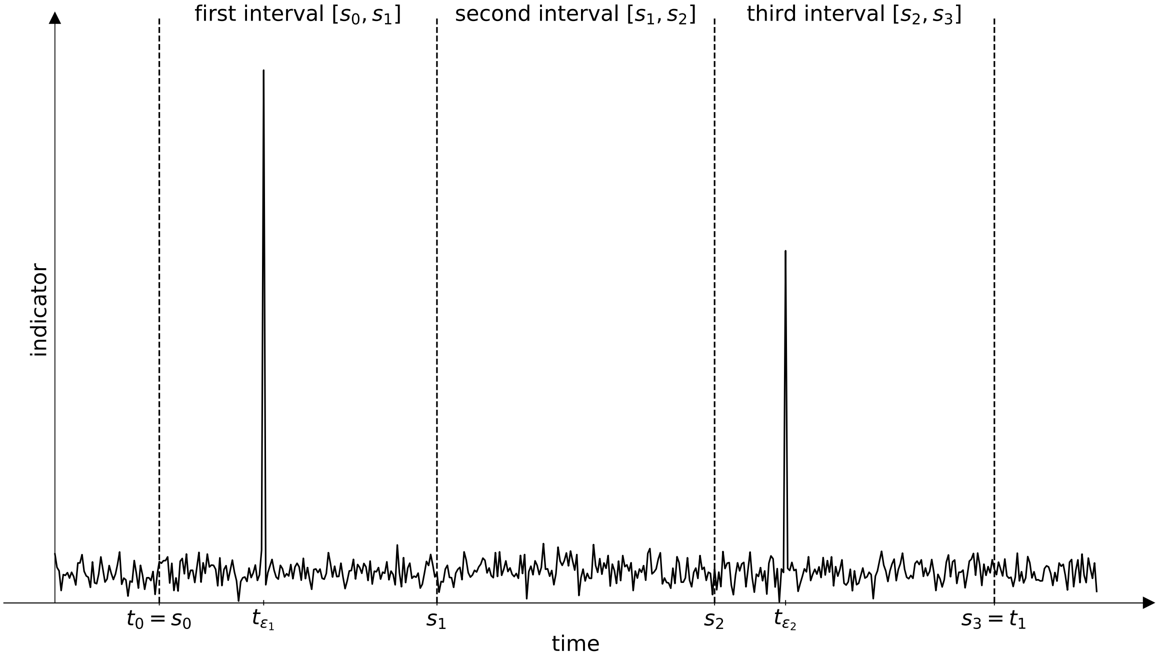

Figure 7: Example of an indicator function

In the example of Figure 7, there are two outliers. Thus, in the first sub-time period, it is expected that the probability of an outlier is very close to one. However, in the third sub-time period, the expectation would be that the conditional probability of an outlier is very low, as the condition is that there was no outlier in the first sub-time period. Nevertheless, in the total time period, there would be a very high probability, as the probability in the first sub-time period is already close to one.

5.2.1 Estimating the unconditional probability

As defined above, an event is a deviation from the normal state and therefore is represented as an outlier of the indicator function. As demonstrated in section 3.1 the Grubbs outlier test is the best fit to detect outlier for our purposes. In most cases, an outlier test is used as a classical statistical hypothesis test to determine if a certain value is an outlier to a significance level. A specification of this significance level needs to be set in advance. However, for our approach, we chose to calculate the highest probability level p at which the hypothesis of no outlier would not be rejected instead of determining a fixed significance level before applying the outlier test. The probability level p can be interpreted as the probability that there is no outlier in the data set. For our purposes, we assume that events are characterized by outliers only on one side, meaning that we assume that very low values are not defined as outliers and we only take very high values into account. However, the proposed approach can be easily transferred to a two-side or a minimum-only approach by minor and self-explanatory changes (Grubbs 1950).

In the case of the unconditional probability $P(O(I, e_i^r, s_0, s_1) \neq \emptyset)$, there are no adjustments necessary and the Grubbs test can be directly applied. The test statistic is defined as

$$G_{s_0,s_1} = \frac{\max\limits_{u_i \in [s_0,s_1]} I(u_i) - \overline{I_{s_0,s_1}}}{\sigma_{s_0,s_1}},$$

with

$$\overline{\mathrm{I}_{s_0,s_1}} = \frac{1}{n} \sum_{i:u_i \in [s_0,s_1]} I(u_i),$$

$$\sigma^2_{s_0,s_1} = \frac{1}{n-1} \sum\nolimits_{i:u_i \in [s_0,s_1]} \{(I(u_i) - \}\overline{I_{s_0,s_1}})^2,$$

where

$$n = |\{i: u_i \in [s_0, s_1]\}|.$$

To conclude the Grubbs test statistic is the maximum of all indicator function values in the interval $[s_0, s_1]$ minus the mean of the indicator function values divided by the corrected standard deviation of the indicator function values. We assume that outliers only appear on one side. Used as a classical hypothesis test with significance level α, the hypothesis that there is no outlier would be rejected if

$$G_{s_0,s_1} > \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{\frac{\alpha}{n},n-2}}{n-2+t^2_{\frac{\alpha}{n},n-2}}},$$

with $t_{\alpha,N}$ representing the upper critical value of a t-distribution with N degrees of freedom and significance level α. Therefore, to get the probability p_{s_0,s_1} for the existence of an outlier in the interval $[s_0, s_1]$, the equation

$$G_{s_0,s_1} = \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{\frac{p_{s_0,s_1}}{n},n-2}}{n-2+t^2_{\frac{p_{s_0,s_1}}{n},n-2}}}$$

needs to be solved with respect to p_{s_0,s_1}. This equation is non-trivial and therefore only solvable by using numeric methods. As a result, the probability p_{s_0,s_1} states the probability that there is an outlier within this sub-time period

$$P(O(I, e_i^r, s_0, s_1) \neq \emptyset) = p_{s_0,s_1}.$$

5.2.2 Estimating the conditional probabilities

As shown above the computation of the probability $P(O(I,e_i^r,s_0,s_1) \neq \emptyset)$ is straightforward. However, it is not entirely obvious how to treat the conditional probability $P\big(O(I,e_i^r,s_{k-1},s_k|) \neq \emptyset | O\big(I,e_i^r,s_{0,}s_{k-1}\big) = \emptyset\big)$ for an arbitrary $k \in \mathbb{N}$. By assuming independence this would be trivial since in that case, the condition would simply disappear. The problem is that this assumption is very strong and most likely not justifiable for most applications. For this reason, we have to estimate the conditional probabilities. An important observation is that we still need to use the Grubbs test to remain consistent with respect to the first sub-time period. Using other, different outlier tests would bias the results and thus violate the condition in the conditional probability. Therefore, we first note that we should use the maximum value from the interval $[s_{k-1},\ s_k]$ in the Grubbs-Test. This is trivial since we are interested in whether we have an outlier within the interval $[s_{k-1},s_k]$ and therefore can only take the highest value from this specific interval into consideration. Second, we note that we cannot use any values from a prospective time point of s_{k+1}, as we calculate the probability for an outlier in the interval $[s_{k-1},s_k]$, conditioned on the past. Note that from a purely mathematical point of view it would be possible to compute the probability of an outlier in the complete time period beginning at the most recent time interval and then conditioning on the future of the time periods. However, this is rather counterintuitive since it would require using data from the future that was not observable at this time point to calculate the probability of an outlier. For this reason, this approach is disregarded from further consideration. In a further observation, we determine that the rather technical proof of Grubbs test (Thompson 1935) is based on the assumption that the sample mean and the sample standard deviation are drawn from the same data set. Otherwise, the statement of the distribution of the test statistic does not longer hold up, and therefore the equation to calculate the probability p_{s_{k-1},s_k} would no longer produce mathematical correct results. Summarizing all the above the best estimation of the conditional probability $P\big(O(I,e_i^r,s_{k-1},s_k|) \neq \emptyset | O\big(I,e_i^r,s_{0,}\ s_{k-1}\big) = \emptyset\big)$ is apparent. As the probability is conditioned on the past the sample mean and the sample standard deviation must be calculated based on the indicator function in the interval $[s_0,s_{k-1}]$, because this is the only possibility to use information of the past in the Grubbs test, which justifies the condition on the past. Moreover, we also must use the attributes from the interval $[s_{k-1},s_k]$, as otherwise the results would be biased in comparison to the first sub-time period, where we use the attributes from the same sub-time period as well. Therefore, we can define the following test statistic for the interval $[s_{k-1},s_k]$:

$$G_{s_{k-1},s_k} = \frac{\max\limits_{u_i \in [s_{k-1},s_k]} I(u_i) - \overline{I_{s_0,s_k}}}{\sigma_{s_0,s_k}},$$

applying the notation described above.

This leads to the following equation

$$G_{s_{k-1},s_k} = \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{\frac{p_{s_{k-1},s_k}}{n},n-2}}{n-2+t^2_{\frac{p_{s_{k-1},s_k}}{n},n-2}}},$$

which needs to be solved for p_{s_{k-1},s_k}. All in all, we conclude that the best estimation for the probability $P\big(O(I,e_i^r,s_{k-1},s_k) \neq \emptyset \big| O\big(I,e_i^r,s_{0,}s_{k-1}\big) = \emptyset\big)$ is given by $p_{s_k,s_{k+1}}$.

In summary, we can summarize that we can calculate the probability of an event in the interval $[t_0,t_1]$ as

$$P(CUR(e_i^r,t_0,t_1) = true) = 1 - \prod_{k=1}^{l} p_{s_{k-1},s_k}.$$

6 Evaluation

To review whether the in the previous chapter proposed metric is suitable for an application on real-world data in this chapter, we evaluate (E1) the practical applicability and (E2) the effectiveness of our approach for a currency metric. First, we discuss the reasons for selecting the case of Wikipedia and describe the analysed dataset. Then, we show how the approach could be instantiated for this case. Finally, we present the results of its application.

6.1 Case Selection and Dataset

To evaluate (E1) and (E2), the approach was applied to a dataset of articles from the English Wikipedia. As one of the most important websites nowadays[11], Wikipedia contains more than 55 million articles in 319 languages[12] and dialects. The English Wikipedia was chosen as it is the most relevant Wiki of Wikipedia due to its size[13] and visitor statistics[14]. One of the key advantages of Wikipedia is having more than 100.000 active voluntary editors instead of an editor desk. However, as a downside, the lack of quality control needs to be emphasized (Agarwal et al. 2020; Anthony et al. 2009; La Robertie et al. 2015; Ofek and Rokach 2015; Warncke-Wang et al. 2015). While most of the edits are doubled checked, there is only little quality control concerning outdated facts. Though, Wikipedia has its own category for outdated

[11] https://www.alexa.com/topsites

[12] https://en.wikipedia.org/wiki/Wikipedia:About

[13] https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

[14] https://pageviews.toolforge.org/siteviews/?platform=all-access&source=pageviews&agent=user&start=2020-01&end=202012&sites=en.wikipedia.org|de.wikipedia.org|fr.wikipedia.org|es.wikipedia.org|ceb.wikipedia.org|sv.wikipedia.org|it.wikipedia.org

articles[15]. To belong to this category articles are manually tagged with a specific date from when a certain fact from said article becomes outdated. Unfortunately, this method is only applicable if an expiry date is known a priori, which is delusive regarding unplanned real-world events. Even in the cases where an expiry date is known a priori, there is zero quality control in Wikipedia if the tagged date is correct in the corresponding article. If, however, a non-regular real-world event happens that alters or adds facts and no one changes the corresponding Wikipedia article, it will be outdated. If no information is added to the article, subsequent users will receive outdated facts about the searched topic.

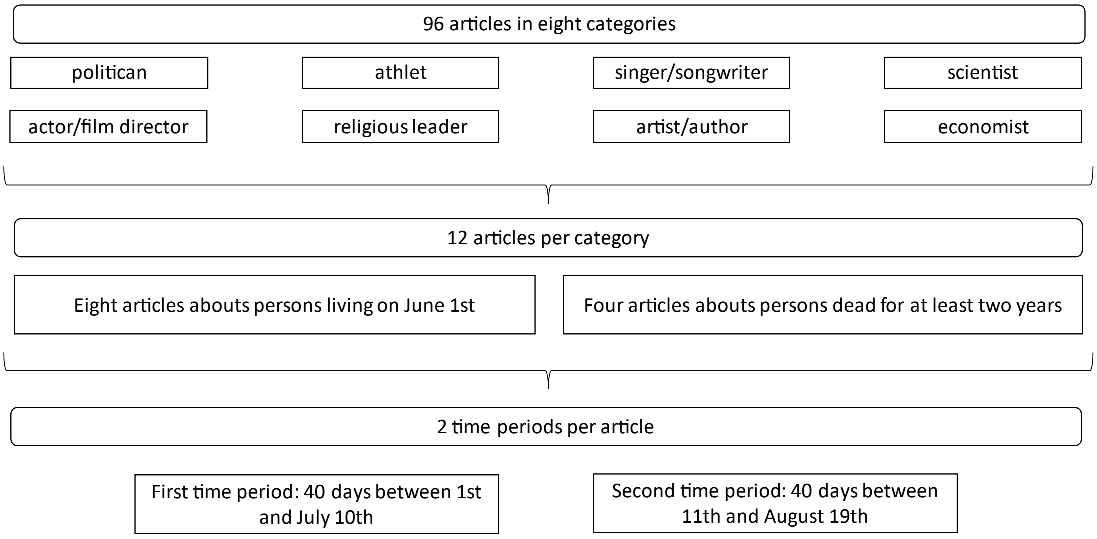
96 articles in eight categories
politican
athlet
singer/songwriter
scientist
actor/film director
religious leader
artist/author
economist
12 articles per category
Eight articles abouts persons living on June 1st
Four articles abouts persons dead for at least two years
2 time periods per article
First time period: 40 days between 1st and July 10th
Second time period: 40 days between 11th and August 19th

Figure 8: Description of the data set of articles

To apply and evaluate our approach with regard to detecting articles where a notable event happened, we selected 96 articles from the English Wikipedia. All of the articles chosen were about persons since articles about single individuals form one of the main categories of Wikipedia. Every fourth article is written about a person and 55% of the most viewed articles in 2019[16] and 2020[17] belong to individuals. Figure 8 explains the selection criteria that we used to select the 96 articles. To define eight categories, we looked at the most clicked articles about persons from 2019 and 2020, as well as the subcategories of the German Wikipedia category “persons by field”[18]. This was done to check if the metric can perform well on a wide range of

[15] https://en.wikipedia.org/wiki/Category:Wikipedia_articles_in_need_of_updating

[16] https://pageviews.toolforge.org/topviews/?project=en.wikipedia.org&platform=all-access&date=2019&excludes=

[17] https://pageviews.toolforge.org/topviews/?project=en.wikipedia.org&platform=all-access&date=2020&excludes=

[18] https://de.wikipedia.org/wiki/Kategorie:Personen_nach_Sachgebiet

different articles about persons who both had and had not have a notable event occur to them recently. For a better evaluation, we chose to apply our metric for each article in two disjunct time intervals, which doubles the number of instances to apply our approach. The first time period was set for the time between 1st June and 10th July 2020, the second time period is the time between 11th July and 19th August 2020. In total, we had a weighted dataset with respect to many different aspects like the occurrence of a notable real-world event, the vital status information (dead or alive), and the background information of said person`s life to check if our metric can perform well in detecting notable events. During the selection process, we manually searched for notable events in one or both time periods, for each article. This research was conducted separately for each article and time period. As a first step, we searched three major online news sites (CCN, New York Times, and BBC) using Google and a customized time range. As the number of search results was limited through the constraints with respect to the time period and website, we were able to read check all headlines of these news sites. If there was no notable event indicated by this search, we followed up our research with a google search without focusing on a certain news site but still a customized time range. Since our sources from the first step are based in the United States and Great Britain, a special focus on individuals from these countries became apparent during our research. To counteract these findings, we especially focused on our second step while investigating persons from countries other than the United States and Great Britain to detect notable events. Our thoroughly conducted research led to an obvious problem of an overwhelming number of search results. For practical reasons, we decided to only check the first ten pages of search results as it appears rather unlikely that no website in our set scope would be reporting about a notable event happening in the life of our targets. By choosing the articles based on the selection criteria shown in Figure 8, we were able to generate labelled data with respect to the existence of a notable event for our approach to evaluate our approach.

6.2 Instantiation of the metric

To measure events, we chose the moving average of the hourly pageviews. The hourly pageviews can be downloaded as a dump from Wikimedia[19]. The indicator based on the pageviews was chosen as Agarwal et al. (2020) and Göbel and Munzert (2018) described that before elections and after other notable events like resignations the pageviews of British and German politicians, who are connected to the corresponding event, have a considerable peak. One possible explanation for this phenomenon is that if a notable event happens a lot of persons use Wikipedia to search for background information on this certain topic. This is exemplarily shown in Figure 9. The left section shows the pageviews of the American football

[19] https://dumps.wikimedia.org/other/pagecounts-ez/merged/

player Patrick Mahomes, who became a minority owner of the baseball team Kansas City Royals on July 28. At the same time, there is a clear peak in the pageviews per hour and therefore in the moving average of the data as well. In the left section, the pageviews per hour of the American scientist Neil deGrasse Tyson are shown. There was no notable event surrounding his name and therefore the moving average is inconspicuous, remaining on a relatively constant level. The high fluctuations seen in the raw data are a result of the different times of a day in the United States, where most of the users of the English Wikipedia are located. These fluctuations made us use a moving average with a length of 48 hours as this is a common method from time series analysis to remove such a seasonality (Winker 2010). 48 hours were chosen as length since the season (a day) has a length of 24 hours, therefore the length must be a multiple of 24. However, with a length of 24, the moving average is very sensitive to small deviations and therefore leading to unstable probability estimations.

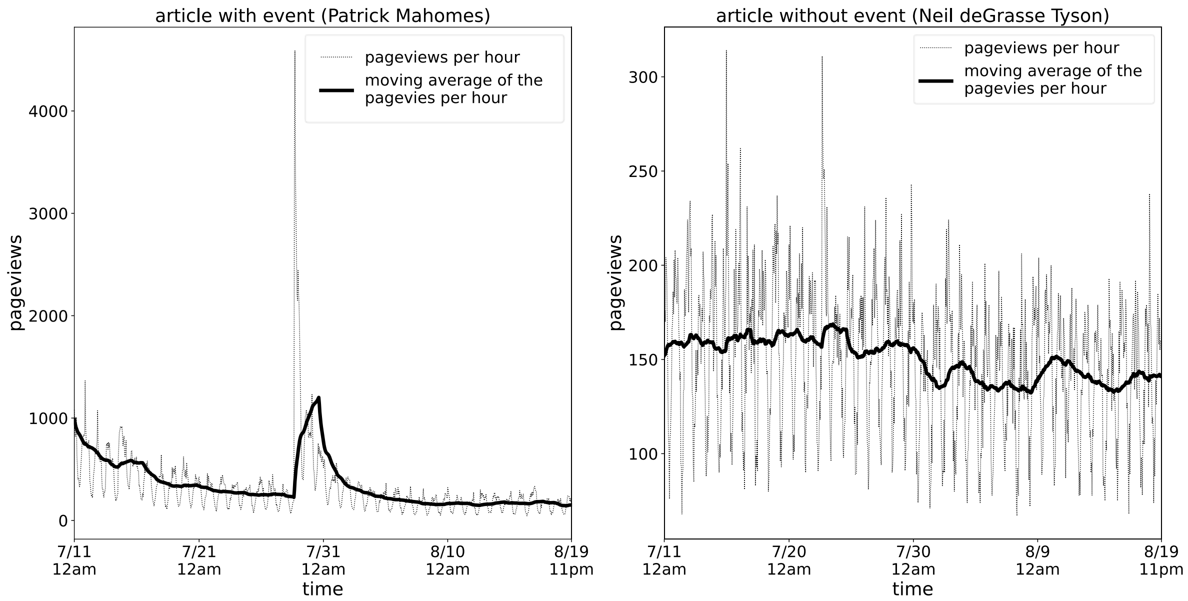

Figure 9: Pageviews as an event indicator

As described in chapter 5 the probability of a notable event is calculated by dividing the complete time period of the pageview moving average into multiple smaller sub-time periods. The probability is calculated for each sub-time period on the condition that there were no outliers in the prior sub-time periods using a one-sided Grubbs outlier test. For our evaluation, we divided each time period into two sub-time periods á 20 days. Consequently, we have at least a sample size of 480 hours in the calculation of the Grubbs statistic. Choosing too short sub-time periods leads to unstable probability estimations.

6.3 Application and results

6.3.1 (E1) Practical applicability

The approach was implemented in Python and applied to all articles in both time periods. After initial instantiation, our approach could be applied in an automated manner without manual configuration. Its practical applicability (E1) is underlined by the low effort required to set up. For each article and each time period, the metric yielded an estimation for the probability of a notable event in the corresponding article and time period. Figure 10 shows a histogram of the estimated probabilities. The relative frequencies are given in five bins according to the estimated event probability. In the majority of the articles, the approach assigned either a very low or a very high event probability in the majority of articles. Such a distribution of estimated event probabilities is favourable as it builds the basis for a clear and comprehensible classification.

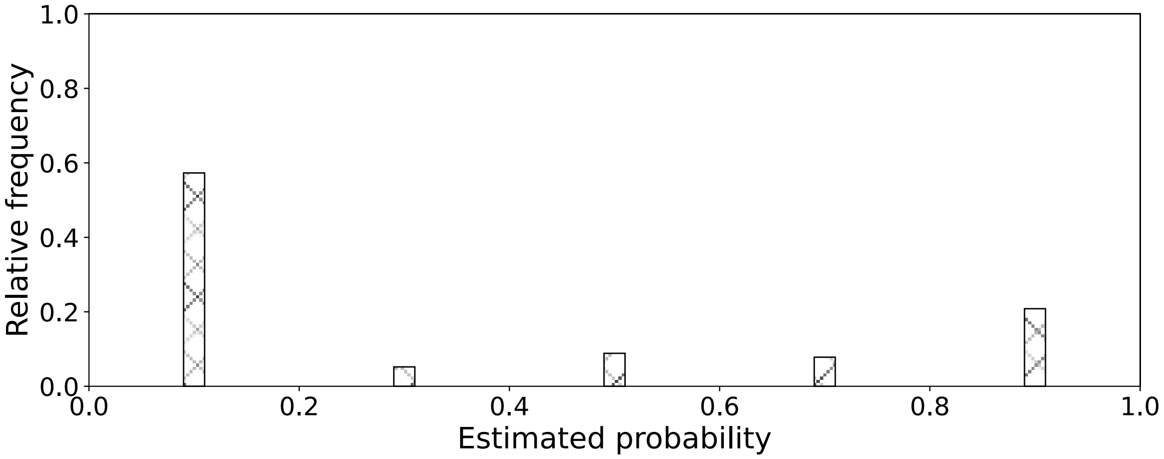

Figure 10: Histogram of estimated event probabilities

Our approach aims to determine event probabilities for an article in a specific time interval, which can be used to classify into two classes where either a notable event was happening or not. Therefore, to evaluate the effectiveness (E2) we first analyse whether the proposed metric is able to provide event probability estimation of high quality (E2.1). Then the effectiveness of our approach with respect to the classification is assessed (E2.2).

6.3.2 (E2.1) Effectiveness with respect to the estimated event probabilities

Regarding (E2), we evaluate the metric with respect to the estimated event probabilities. For an extensive discussion cf. chapter 3. To be a useful metric, it needs to be ensured that the estimated probabilities correspond to the actually observed relative frequencies, which can be assessed in terms of reliability (Hoerl and Fallin 1974; Murphy and Winkler 1977; Murphy and Winkler 1987; Sanders 1963). In our context reliability conveys that the relative frequencies of real-world events within an interval must be approximately equal to the mean of the estimated event probabilities. A common way to evaluate reliability is the reliability curve (Bröcker and Smith 2007). To calculate the points of this curve, the data is ordered into bins according to the estimated event probability. Afterward, the mean of the estimated event probability ("mean estimated probability"), and the actual relative frequency of real-world events ("fraction of positives"), are calculated and plotted for each bin separately. A perfectly reliable estimation would be characterized by all points of the reliability curve lying on the diagonal. On the left side of Figure 11, the reliability curve for our approach is shown. To obtain a sufficient number of test cases in each bin, the number of bins was set to five. The results show that our approach assigns reliable event probabilities to the articles in the different time periods as the curve follows the diagonal rather closely. However, our approach seems to slightly overestimate the probabilities in the realm of high probabilities above 50%.

Based on the event probabilities estimated by our approach, events can be distinguished from non-events. Thus, to evaluate this aspect, we determined the discrimination of the estimated duplicate probabilities. The discrimination was assessed in terms of the AUC below the ROC curve (Hanley and McNeil, 1982). The ROC curve is calculated by plotting the true positive rate of a classification based on the estimated duplicate probabilities against the false positive rate when the classification threshold is varied. The ROC curve is given in the right section of Figure 11. The ROC curve is close to the curve of perfect discrimination. The area under the ROC curve amounts to 88.46%, which is considered as excellent discrimination in the literature (Hosmer et al.). Overall, these results support that the probabilities provided by our approach are able to discriminate between events and non-events. Further, they motivate the classification of pairs of records into events and non-events based on this approach, which the following is focused on.

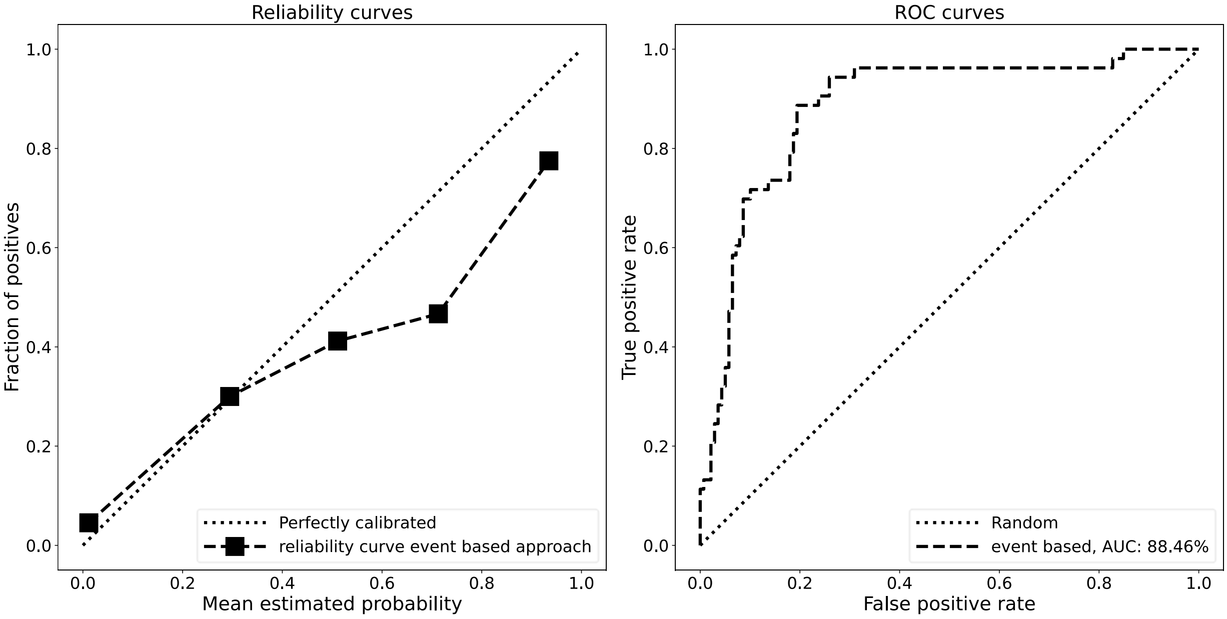

Figure 11: Reliability curve and ROC curve for the probability-based metric

6.3.3 (E2.2) Effectiveness with respect to classification into events and non-events

Chapter 3 contains a detailed discussion on the evaluation of a classifier based on a probability-based metric. In this section, we only briefly discuss the theoretical backgrounds. To classify into events and non-events the articles with probability over an a priori determined threshold in the corresponding time period were classified as articles with a notable event and vice versa. We chose two different thresholds. For the first one, we picked 50%, which is very intuitive. As the second threshold, we chose the optimal cut-off point based on Yourden's J-statistic (Youden 1950; Schisterman et al. 2005). The optimal cut-off value for our approach is 34.79%. To assess the quality of the classification into events and non-events, the performance measures accuracy, precision, recall, and F-measure (F_1) are provided in Table 6 for both thresholds. F-measure combines precision and recall and is defined as their harmonic mean. On the given dataset, our approach provides very promising results. For instance, using the optimal cut-off value of 34.79% our approach was able to identify 88.67% of all articles which needed to be updated (recall event). Even given the natural threshold of 50%, our approach was still able to classify 73.58% of the articles correctly that had a notable event. This shows that our approach is capable of identifying notable events. Furthermore, in both cases, the classification leads to a high accuracy of 82.81% for the optimal cut-off value and 79.69% for the 50% threshold. As already stated above our approach slightly overestimates the probability in the realm of high probabilities over 50%, which leads to a precision for events of 63.51% (optimal cut-off value) respective 60.94% (50% threshold).

	Accuracy	Precision (event)	Recall (event)	F_1 (event)	Precision (no event)	Recall (no event)	F_1 (no event)
Threshold 50%	79.69%	60.94%	73.58%	66.67%	89.06%	82.01%	85.39%
Optimal cut-off point	82.81%	63.51%	88.67%	74.01%	94.92%	80.58%	87.16%

Table 6: Performance measures for classification into events and no-events

To conclude our approach showed promising results regarding (E2.2) the performance of the classifier based on the estimated event probabilities. It also applies to effectiveness due to the estimated event probabilities.

7 Conclusion

After presenting the results of the proposed metric on a real-world data set in the previous chapter, in the following final chapter, we want to discuss the results of our approach, show limitations and outline further work and research to present an extensive conclusion of our proposed probability-based event metric for measuring the currency.

7.1 Discussion of the results

As shown in the reliability plot in the previous section (cf. Figure 11) our approach gives very reliable predictions especially in the interval $[0, 0.5]$ the line is very close to the diagonal line of a perfectly calibrated probability estimator. However, our approach seems to slightly overestimate the probabilities in the realm of high probabilities above 50%. Several reasons factor into these overestimations. Most importantly, especially striking in articles, which are averagely not categorized popular, even comparatively small increases in the pageviews over a period of a few hours can result in a high probability of an outlier. Such small distinctive abnormalities recorded over few hours can be the consequence of randomness, but also arise from interviews or news articles even in small newspapers and blogs. A reason for the high probability estimations is that a theoretical requirement of the Grubbs test is that the data are approximately normally distributed, which probably does not apply to all articles. Moreover, it is possible that peaks due to events, which occurred in a previous time period are used in the moving average calculation, which obviously can bias the computations. On the other side, it can help to identify notable events that occurred shortly before the reviewed time period. Although this was not set as the initial goal of the approach it can be helpful in a practical application of this approach. Finally, a Wiki article may achieve high public interest due to an event in a related real-world entity. For instance, a film remake results in high interest in the film director of the original movie production or even persons with similar names that get mixed up. Nevertheless, it can be stated that our proposed approach showed very promising results. This is further emphasised by the very high-quality ROC curve our approach was able to achieve. This proves that the proposed probability-based metric can be used as a classifier as well. Thus, the results as a classifier were very promising as well (cf. Table 6). Using the optimal cut-off point determined using Youden's J statistic we achieved a high accuracy of 83%. More importantly, 89% (recall events) of the articles with a notable event were classified correctly. Even as the problems described above lead to a mediocre precision in event class, we were still able to classify 81% (recall no events) of the articles without notable events correctly. Overall, the F_1 measure, which combines both the precision and the recall, is high for both classes with 74% (articles with a notable event) and 87% (articles without notable event). Even using the natural threshold of 50% still 74% (recall events) containing a notable event were classified correctly with a still high accuracy of 80%. Using our approach can further improve the quality of Wiki articles especially in the context of updated articles. For instance, our approach detected an event in the article about Pope Benedict XVI in early August. At this time, it was publicly announced that his health condition got worsen[20]. However, in Wikipedia, the article was only updated in January 2021. The article about the English street artist Banksy was not updated until February 2021. Even though Banksy drew a lot of attention with his painting in the London tube[21] in July 2020. This event was detected by our approach as well.

Compared to other existing approaches such as Tran and Cao (2013) our approach shows better results especially since our approach focuses on a more general setting since Tran and Cao (2013) limited their approach to infoboxes. Nevertheless, our approach is able to achieve a higher accuracy, independently whether using the optimal cut-off value or the 50% cut-off value. Moreover using the optimal cut-off point the proposed approach is also able to achieve a higher recall than the approach of Tran and Cao (2013). Unfortunately, it is hardly possible to compare our probability-based metric to the approach of Stvilia et al. (2005b). In particular, their definition of currency, which is characterized as the time since the last update, seems compared to our approach very difficult to interpret. Using our approach has therefore the distinct advantage that a probability can be understood very easily.

7.2 Limitations and future research

Besides the great accomplishments, there are limitations to our approach, which need to be considered. At first, it needs to be recognized that our approach can only detect notable events. implying the are some update-relevant events this approach might oversee. Such events include all incidents that are not of public interest like for example updates in statistics like population numbers. Thus, future research could explore whether it is possible to develop an approach to also detect update relevant events without major interest of the Wiki users or if it is even possible to find a method that integrates the best of both worlds and to develop a comprehensive approach.

The starting point of such an approach could be the work of Heinrich and Klier (2015) to model events, where average shelf lives are known, or Pernici and Scannapieco (2003) in case of known expiry dates. Such an approach could take different metadata into account such as the Wikipedia intern categories an article belongs to. Figure 12 is showing an encouraging example that this approach may work is shown. In this plot, the empirical cdf of the number of days between two edits for two different categories (of the German Wikipedia) persons born in the 13th century and persons born after 1960 is presented. This can only be a starting point

[20] https://thetablet.org/vatican-confirms-pope-benedict-is-ill-but-says-condition-not-serious/

[21] https://www.bbc.com/news/uk-england-london-53407715

since not all edits of Wikipedia articles are made to change outdated facts but also to include additional facts or sources. Nevertheless, the results are promising since it is evident that articles about individuals born after 1960 and thus with a high probability of still being alive are updated more frequently than articles about individuals that are categorized dead for over 600 years.

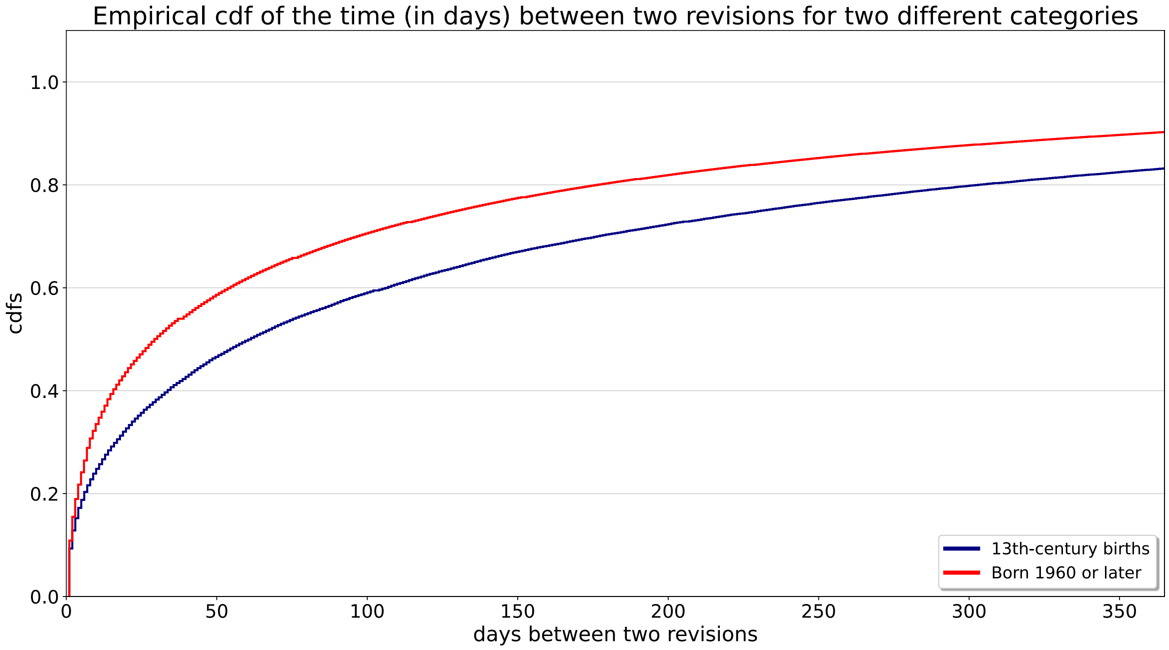

Figure 12: Empirical cdf of the days between an update for two different categories of the German Wikipedia

Naturally, persons, who are still alive have an increased chance of being involved in an occurrence leading to a relevant update to their article – especially compared to an individual who has been dead for centuries. Future research could focus on the development of a method that takes different categories an article belongs to into account to derive an estimated update frequency cdf. One possible way could be to determining the k most similar articles based on the categories and calculating an update frequency cdf based on these articles. To gain more precise results one could try to expand the approach beyond categories and take more metadata of an article into account like the used templates, internal links and used pictures. Furthermore, it could be very helpful evolving an approach that can distinguish between edits made due to change or remove outdated facts and edits that are made for different reasons in order to achieve more precise update frequency cdfs.

Moreover, our approach is not able to detect the type of notable event that occurred. In practice, this means that there is still a manual editor necessary to update the corresponding events. This however is a very common limitation of currency metrics. Therefore, further research could try to develop an approach that automatically detects which kind of event took place and may even automatically edit the article. Such an approach would probably be using NLP and automatically searching through major news sites and/or major search engines.

A more general limitation of all approaches for assessing the data quality of Wiki articles is that is hardly possible to define the scope of the content of a Wiki article. This issue refers to currency but also other data quality dimensions such as completeness and accuracy. Nevertheless, we will discuss it in this section with a focus on the currency. Most Wikis define a quality standard and describe what criteria distinguish superlative articles from others. For instance, Wikipedia describes that featured articles must be comprehensive, meaning that they do not neglect any major facts[22]. However, it is not specified how exactly a major fact is defined. For that reason, compiling a definition, which would fit all articles in every given situation, is practically impossible. Due to this fact, assessing the currency of Wiki articles partly a subjective decision. Besides significant events that are undoubtedly resulting in an update like the death of a person, there are also incidents where the deciding factor for or against update relevance can be weighted very subjectively. For instance, who is deciding if it is update relevant when a businessman donates for a charity supporting poor children or if updating an enterprise Wiki after a company's CEO became a parent is relevant. It is not apparent what attributes would lead to a categorisation as update relevant. Using our approach this implies that it may occur that events are detected, that are not evident whether they have update relevance. Nevertheless, by detecting such events our approach builds awareness of editors, starting a discussion if an event is update relevant. Thus, future research could try to develop criteria for which events and facts can be categorized as major facts and which events and facts can be ignored.

.

[22] https://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria

8 References

Adikaram, K. K. L. B., Hussein, M. A., Effenberger, M. and Becker, T. 2015. "Data Transformation Technique to Improve the Outlier Detection Power of Grubbs' Test for Data Expected to Follow Linear Relation," *Journal of Applied Mathematics* 1–9.

Adler, B. T. and Alfaro, L. 2007. "A Content-Driven Reputation System for the Wikipedia," in *Proceedings of the 16th International Conference on World Wide Web,* Banff, Canada.

Agarwal, P., Redi, M., Sastry, N., Wood, E. and Blick, A. 2020. "Wikipedia and Westminster: Quality and dynamics of Wikipedia pages about UK politicians," in *Proceedings of the 31st ACM Conference on Hypertext and Social Media,* Orlando, FL.

Aitamurto, T., Landemore, H. and Saldivar Galli, J. 2017. "Unmasking the crowd: participants' motivation factors, expectations, and profile in a crowdsourced law reform," *Information, Communication & Society* (20:8), 1239–1260.

Anderka, M. and Stein, B. 2012a. "A breakdown of quality flaws in Wikipedia," in *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality,* Lyon, France.

Anderka, M., Stein, B. and Lipka, N. 2012b. "Predicting quality flaws in user-generated content: the case of wikipedia," in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval,* Portland, OR.

Anthony, D., Smith, S. W. and Williamson, T. 2009. "Reputation and Reliability in Collective Goods," *Rationality and Society* (21:3), 283–306.

Arolfo, F. and Vaisman, A. 2018. "Data Quality in a Big Data Context," in *Proceedings of the 22nd European Conference on Advances in Databases and Information Systems,* Budapest, Hungary.

Aslam, M. 2020. "Introducing Grubbs's test for detecting outliers under neutrosophic statistics – An application to medical data," *Journal of King Saud University - Science* (32:6), 2696–2700.

Ballou, D., Wang, R., Pazer, H. and Tayi, G. K. 1998. "Modeling Information Manufacturing Systems to Determine Information Product Quality," *Management Science* (44:4), 462–484.

Batini, C., Daniele, B., Federico, C. and Simone, G. 2011. "A Data Quality Methodology for Heterogeneous Data," *International Journal of Database Management Systems* (3:1), 60–79.

Batini, C., Palmonari, M. and Viscusi, G. 2014. "Opening the Closed World: A Survey of Information Quality Research in the Wild," In: Floridi, L. and Phyllis Illari (eds.) in *The Philosophy of Information Quality* Cham, Switzerland: Springer, pp. 43–73.

Batini, C. and M. Scannapieco. 2016. *Data and information Quality:* Dimensions, Principles and Technique, Cham, Switzerland: Springer International Publishing.

Benkhaled, H. N. and Berrabah, D. 2019. "Data Quality Management For Data Warehouse Systems: State Of The Art," in *The Proceedings of JERI'2019,* Saïda, Algeria.

Bhatti, Z. A., Baile, S. and Yasin, H. M. 2018. "Assessing enterprise wiki success from the perspective of end-users: an empirical approach," *Behaviour & Information Technology* (37:12), 1177–1193.

Blake, R. H. and Mangiameli, P. 2009. "Evaluating the semantic and representational consistency of interconnected structured and unstructured data," in *Proceedings of the 15th Americas Conference on Information Systems,* San Francisco, CA.

Blumenstock, J. E. 2008a. "Automatically assessing the quality of Wikipedia articles," *UCB iSchool Report* (2008-021).

Blumenstock, J. E. 2008b. "Size matters: word count as a measure of quality on wikipedia," in *Proceedings of the 17th international conference on World Wide Web,* Bejing, China.

Bovee, M., Srivastava, R. P. and Mak, B. 2003. "A conceptual framework and belief-function approach to assessing overall information quality," *International Journal of Intelligent Systems* (18:1), 51–74.

Brandes, U., Kenis, P., Lerner, J. and van Raaij, D. 2009. "Network analysis of collaboration structure in Wikipedia," in *Proceedings of the 18th international conference on World wide web,* Madrid, Spain.

Bröcker, J. and Smith, L. A. 2007. "Increasing the Reliability of Reliability Diagrams," *Weather and Forecasting* (22:3), 651–661.

Bykau, S., Korn, F., Srivastava, D. and Velegrakis, Y. 2015. "Fine-grained controversy detection in Wikipedia," in *Proceedings of the 2015 IEEE 31st International Conference on Data Engineering (ICDE),* Seoul, Korea.

Caballero, I., Serrano, M. and Piattini, M. 2014. "A Data Quality in Use Model for Big Data," in *Proceedings of the 33rd Internation Conference on Conceptual Modeling,* Atlanta, GA.

Chavuenet, W. 1871. *A Manual of Spherical and Practical Astronomy*.

Cheng, L. T. E., Zheng, J., Savova, G. K. and Erickson, B. J. 2010. "Discerning tumor status from unstructured MRI reports-completeness of information in existing reports and utility of automated natural language processing," *Journal of Digital Imaging* (23:2), 119–132.

Cho, J. and Garcia-Molina, H. 2003. "Effective page refresh policies for Web crawlers," *ACM Transactions on Database Systems* (28:4), 390–426.

Collinson, P. 1998. "Of bombers, radiologists, and cardiologists: time to ROC," *Heart* (80:3), 215–217.

Cykana, P., Paul, A. and Stern, M. 1996. "DoD Guidelines on Data Quality Management," in *Proceedings of the 1996 Conference on Information Quality,* Cambridge, MA.

Dalip, D. H., Gonçalves, M. A., Cristo, M. and Calado, P. 2009. "Automatic Quality Assessment of Content Created Collaboratively by Web Communities: A Case Study of

Wikipedia," in *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries,* Austin, TX.

Dalip, D. H., Gonçalves, M. A., Cristo, M. and Calado, P. 2017. "A general multiview framework for assessing the quality of collaboratively created content on web 2.0," *Journal of the Association for Information Science and Technology* (68:2), 286–308.

Dalip, D. H., Lima, H., Goncalves, M. A., Cristo, M. and Calado, P. 2014. "Quality assessment of collaborative content with minimal information," in *Proceedings of the 2014 IEEE/ACM Joint Conference on Digital Libraries (JCDL),* London, UK.

Dang, Q. V. and Ignat, C.-L. 2016a. "Quality assessment of wikipedia articles: a deep learning approach," *ACM SIGWEB Newsletter* (16:Autumn), 1–6.

Dang, Q. V. and Ignat, C.-L. 2016b. "Quality Assessment of Wikipedia Articles without Feature Engineering," in *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries,* Newark, NJ.

Dang, Q.-V. and Ignat, C.-L. 2016c. "Measuring Quality of Collaboratively Edited Documents: The Case of Wikipedia," in *Proceedings of the 2016 IEEE 2nd International Conference on Collaboration and Internet Computing,* Pittsburgh, PY.

Dang, Q.-V. and Ignat, C.-L. 2017. "An end-to-end learning solution for assessing the quality of Wikipedia articles," in *Proceedings of the 13th International Symposium on Open Collaboration,* Berlin, Germany.

Dardis, C. 2004. "Peirce's criterion for the rejection of non-normal outliers: defining the range of applicability," *Journal of Statistical Software* (10), 1–8.

Dean, R. B. and Dixon, W. J. 1951. "Simplified Statistics for Small Numbers of Observations," *Analytical Chemistry* (23:4), 636–638.

Dennis, J. E. and R. B. Schnabel. 1996. *Numerical methods for unconstrained optimization and nonlinear equations.* Classics in applied Mathematics Vol. 16, Philadelphia: SIAM.

Di Sciascio, C., Strohmaier, D., Errecalde, M. and Veas, E. 2017. "WikiLyzer: interactive information quality assessment in Wikipedia," in *Proceedings of the 22nd International Conference on Intelligent User Interfaces,* Limassol, Cyprus.

Di Sciascio, C., Strohmaier, D., Errecalde, M. and Veas, E. 2019. "Interactive Quality Analytics of User-generated Content: An Integrated Toolkit for the Case of Wikipedia," *ACM Transactions on Interactive Intelligent Systems (TiiS)* (9:2-3), 1–42.

Dufty, D., Bérard, H., Lefranc, S. and M. Signore. 2014. *A suggested Framework for the Quality of Big Data: Deliverables of the UNECE Big Data Quality Task Team*, Project Deliverable Big Data Quality Framework: UNECE/HLG.

Emigh, W. and Herring, S. C. 2005. "Collaborative Authoring on the Web: A Genre Analysis of Online Encyclopedias," in *Proceedings of the thirty-eighth annual Hawaii international conference on system sciences,* Big Island, HI.

English, L. P. 1999. *Improving data warehouse and business information quality: methods for reducing costs and increasing profits*, New York, NY: John Wiley & Sons.

Eppler, M. J. and Muenzenmayer, P. 2002. "Measuring Information Quality in the Web Context: A Survey of State-of-the-Art Instruments and an Application Methodology," in *Proceedings of the Seventh International Conference on Information Quality,* Cambridge, MA.

Even, A., Shankaranarayanan, G. and Berger, P. D. 2010. "Evaluating a model for cost-effective data quality management in a real-world CRM setting," *Decision Support Systems* (50:1), 152–163.

Fawcett, T. 2006. "An introduction to ROC analysis," *Pattern Recognition Letters* (27:8), 861–874.

Feller, W. 2008. *An introduction to probability theory and its applications, vol 2*, New York, NY: John Wiley & Sons.

Firmani, D., Mecella, M., Scannapieco, M. and Batini, C. 2016. "On the meaningfulness of “big data quality”," *Data Science and Engineering* (1.1), 6–20.

Fominykh, M., Prasolova-Førland, E., Divitini, M. and Petersen, S. A. 2016. "Boundary objects in collaborative work and learning," *Information Systems Frontiers* (18:1), 85–102.

Gardyn, E. 1997. "A Data Quality Handbook for a Data Warehouse," in *Conference of Information Quality 1997,* Cambridge, MA.

Ge, M. and Helfert, M. 2007. "A review of information quality research—develop a research agenda," in *Proceedings of the 12th International Conference on Information Quality,* Cambridge, MA.

Göbel, S. and Munzert, S. 2018. "Political Advertising on the Wikipedia Marketplace of Information," *Social Science Computer Review* (36:2), 157–175.

Gould, B. A. 1855. "On Peirce's Criterion for the Rejection of Doubtful Observations, with tables for facilitating its application," *The Astronomical Journal* (4), 81–87.

Grattan-Guiness, I. and Ledermann, W. 1994. "Matrix Theory," In: Grattan-Guiness, I. (ed.) in *Companion Encyclopedia of the History and Philosophy of the Mathematical Sciences* London, UK: Routledge, pp. 775–786.

Grubbs, F. E. 1950. "Sample Criteria for Testing Outlying Observations," *The Annals of Mathematical Statistics* (21:1), 27–58.

Grubbs, F. E. 1969. "Procedures for Detecting Outlying Observations in Samples," *Technometrics* (11:1), 1–21.

Halfaker, A. 2017. "Interpolating Quality Dynamics in Wikipedia and Demonstrating the Keilana Effect," in *Proceedings of the 13th International Symposium on Open Collaboration,* Berlin, Germany.

Hand, D. J. 2009. "Measuring classifier performance: a coherent alternative to the area under the ROC curve," *Machine Learning* (77:1), 103–123.

Hao, S., Chai, C., Li, G., Tang, N., Wang, N. and Yu, X. 2020. "Outdated Fact Detection in Knowledge Bases," in *Proceedings of the 2020 IEEE 36th International Conference on Data Engineering,* Dallas, TX.

Hatcher-Gallop, R., Fazal, Z. and Oluseyi, M. 2009. "Quest for excellence in a wiki-based world," in *Proceedings of the IEEE International Professional Communication Conference, 2009,* Honolulu, HI.

He, W. and Yang, L. 2016. "Using wikis in team collaboration: A media capability perspective," *Information & Management* (53:7), 846–856.

Heidrich, B., Kása, R., Shu, W. and Chandler, N. 2015. "Worlds Apart But Not Alone: How Wiki Technologies Influence Productivity and Decision-Making in Student Groups," *Decision Sciences Journal of Innovative Education* (13:2), 221–246.

Heinrich, B., Hristova, D., Klier, M., Schiller, A. and Szubartowicz, M. 2018a. "Requirements for data quality metrics," *Journal of Data and Information Quality (JDIQ)* (9:2), 1–32.

Heinrich, B., Kaiser, M. and Klier, M. 2007. "How to measure Data Quality? A Metric-based Approach," in *Proceedings of the 28th International Conference on Information Systems (ICIS). Montreal, Queen's University,* Montreal, Canada.

Heinrich, B. and Klier, M. 2011. "Assessing data currency—a probabilistic approach," *Journal of Information Science* (37:1), 86–100.

Heinrich, B. and Klier, M. 2015. "Metric-based data quality assessment — Developing and evaluating a probability-based currency metric," *Decision Support Systems* (72), 82–96.

Heinrich, B., Klier, M. and Kaiser, M. 2009. "A procedure to develop metrics for currency and its application in CRM," *Journal of Data and Information Quality (JDIQ)* (1:1), 1–28.

Heinrich, B., Klier, M., Obermeier, A. A. and Schiller, A. 2018b. "Event-driven duplicate detection: a probability-based approach," in *ECIS 2018 Proceedings,* Portsmouth, UK.

Hodge, V. and Austin, J. 2004. "A Survey of Outlier Detection Methodologies," *02692821* (22:2), 85–126.

Hoerl, A. E. and Fallin, H. K. 1974. "Reliability of Subjective Evaluations in a High Incentive Situation," *Journal of the Royal Statistical Society. Series A (General)* (137:2), 227.

Hosmer, D. W., Lemeshow, S. and R. X. Sturdivant. 2013. *Applied logistic regression.* Wiley series in probability and statistics. Third edition, Hoboken, New Jersey: Wiley.

Huang, H., Stvilia, B., Jörgensen, C. and Bass, H. W. 2012. "Prioritization of data quality dimensions and skills requirements in genome annotation work," *Journal of the American Society for Information Science and Technology* (63:1), 195–207.

Immonen, A., Paakkonen, P. and Ovaska, E. 2015. "Evaluating the Quality of Social Media Data in Big Data Architecture," *IEEE Access* (3), 2028–2043.

Jain, R. B. 2010. "A recursive version of Grubbs' test for detecting multiple outliers in environmental and chemical data," *Clinical Biochemistry* (43:12), 1030–1033.

Javanmardi, S., Lopes, C. and Baldi, P. 2010. "Modeling user reputation in wikis," *Statistical Analysis and Data Mining: The ASA Data Science Journal* (3:2), 126–139.

Jemielniak, D. and Wilamowski, M. 2017. "Cultural diversity of quality of information on Wikipedias," *Journal of the Association for Information Science and Technology* (68:10), 2460–2470.

Jhandir, M. Z., Tenvir, A., On, B.-W., Lee, I. and Choi, G. S. 2017. "Controversy detection in Wikipedia using semantic dissimilarity," *Information Sciences* (418-419), 581–600.

Kiefer, C. 2016. "Assessing the Quality of Unstructured Data: An Initial Overview," in *LWDA 2016 Proceedings,* Potsdam.

Kiefer, C. 2019. "Quality indicators for text data," in *18th symposium of "Database systems for Business, Technology and Web",* Rostock, Germany.

Kiniti, S. and Standing, C. 2013. "Wikis as knowledge management systems: issues and challenges," *Journal of Systems and Information Technology* (15:2), 189–201.

Kittur, A. and 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,* San Diego, CA.

Klobas, J. E. 2006. *Wikis:* Tools for information work and collaboration. Chandos information professional series, Oxford: Chandos.

Kumar, V. and Thareja, R. 2013. "A Simplified Approach for Quality Management in Data Warehouse," *International Journal of Data Mining & Knowledge Management Process* (3:5).

La Robertie, B. de, Pitarch, Y. and Teste, O. 2015. "Measuring Article Quality in Wikipedia using the Collaboration Network," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015,* Paris, France.

Laranjeiro, N., Soydemir, S. N. and Bernardino, J. 2015. "A Survey on Data Quality: Classifying Poor Data," in *Proceedings of the 2015 IEEE 21st Pacific Rim International Symposium on Dependable Computing,* Zhangjiajie, China.

Lee, Y. W., Strong, D. M., Kahn, B. K. and Wang, R. Y. 2002. "AIMQ: a methodology for information quality assessment," *Information & Management* (40:2), 133–146.

Li, F., Nastic, S. and Dustdar, S. 2012. "Data Quality Observation in Pervasive Environments," in *Proceedings of the 15th IEEE International Conference on Computational Science and Engineering,* Paphos, Cyprus.

Li, X., Luo, Z., Pang, K. and Wang, T. 2013. "A Lifecycle Analysis of the Revision Behavior of Featured Articles on Wikipedia," in *2013 International Conference on Information Science and Cloud Computing Companion (ISCC-C),* Guangzhou, China.

Liaw, S. T., Rahimi, A., Ray, P., Taggart, J., Dennis, S., Lusignan, S. de, Jalaludin, B., Yeo, A. E. T. and Talaei-Khoei, A. 2013. "Towards an ontology for data quality in integrated chronic disease management: a realist review of the literature," *International Journal of Medical Informatics* (82:1), 10–24.

Lih, A. 2004. "Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource," *Nature* (3.1), 1–31.

Lim, E., Vuong, B., Lauw, H. W. and Sun, A. 2006. "Measuring Qualities of Articles Contributed by Online Communities," in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence,* Hong Kong, China.

Limb, B. J., Work, D. G., Hodson, J. and Smith, B. L. 2017. "The Inefficacy of Chauvenet's Criterion for Elimination of Data Points," *Journal of Fluids Engineering* (139:5), 054501 1-3.

Lobo, J. M., Jiménez-Valverde, A. and Real, R. 2008. "AUC: a misleading measure of the performance of predictive distribution models," *Global Ecology and Biogeography* (17:2), 145–151.

Loshin, D. 2011. *The Practitioner's Guide to Data Quality Improvement*, Burlington, MA: Morgan Kaufmann Publishers.

Maddala, G. S. and K. Lahiri. 1992. *Introduction to econometrics*, New York, NY: Macmillan.

Matschke, C., Moskaliuk, J. and Kimmerle, J. 2013. "The impact of group membership on collaborative learning with wikis," *Cyberpsychology, behavior and social networking* (16:2), 127–131.

McAfee, A. P. 2006. "Enterprise 2.0: The dawn of emergent collaboration," *MIT Sloan Management Review* (47:3), 21–28.

Merzbach, U. C. and C. B. Boyer. 2011. *A history of mathematics.* Third edition, Hoboken, NJ: John Wiley and Sons.

Mesgari, M., Okoli, C., Mehdi, M., Nielsen, F. Å. and Lanamäki, A. 2015. ""The sum of all human knowledge": A systematic review of scholarly research on the content of Wikipedia," *Journal of the Association for Information Science and Technology* (66:2), 219–245.

Moraga, C., Moraga, M. Á., Calero, C. and Caro, A. 2009. "SQuaRE-Aligned Data Quality Model for Web Portals," in *2009 9th International Conference on Quality Software,* Jeju, Korea.

Morgan, J. T., Gilbert, M., McDonald, D. W. and Zachry, M. 2013. "Project talk: Coordination work and group membership in WikiProjects," in *Proceedings of the 9th International Symposium on Open Collaboration,* Hong Kong, China.

Murphy, A. H. 1973. "A New Vector Partition of the Probability Score," *Journal of Applied Meteorology* (12:4), 595–600.

Murphy, A. H. and Winkler, R. L. 1977. "Reliability of Subjective Probability Forecasts of Precipitation and Temperature," *Applied Statistics* (26:1), 41.

Murphy, A. H. and Winkler, R. L. 1987. "A General Framework for Forecast Verification," *Monthly Weather Review* (115:7), 1330–1338.

Naumann, F. and Rolker, C. 2000. "Assessment methods for information quality criteria," in *Proceedings of the 2000 Conference on Information Quality,* Cambridge, MA.

Nelson, R. R., Wixom, B. H. and Todd, P. R. 2005. "Antecedents of Information and System Quality: An Empirical Examination Within the Context of Data Warehousing," *Journal of Management Information Systems* (21:4), 199–235.

Ofek, N. and Rokach, L. 2015. "A classifier to determine which Wikipedia biographies will be accepted," *Journal of the Association for Information Science and Technology* (66:1), 213–218.

Orr, K. 1998. "Data quality and systems theory," *Communications of the ACM* (41:2), 66–71.

Parssian, A., Sarkar, S. and Jacob, V. S. 2004. "Assessing Data Quality for Information Products: Impact of Selection, Projection, and Cartesian Product," *Management Science* (50:7), 967–982.

Peacock, T., Fellows, G. and Eustace, K. 2007. "The quality and trust of wiki content in a learning community," in *Proceedings of the ascilite 2007,* Singapore, Singapore.

Pearson, E. S. and Sekar, C. C. 1936. "The Efficiency of Statistical Tools and A Criterion for the Rejection of Outlying Observations," *Biometrika* (28:3/4), 308–320.

Pei Lyn Grace, T. 2009. "Wikis as a knowledge management tool," *Journal of Knowledge Management* (13:4), 64–74.

Peirce, B. 1852. "Criterion for the rejection of doubtful observations," *The Astronomical Journal* (2), 161–163.

Pernici, B. and Scannapieco, M. 2003. "Data Quality in Web Information Systems," *Journal on Data Semantics I* 48–68.

Pfaff, C. C. and Hasan, H. M. 2006. "Overcoming organisational resistance to using Wiki technology for Knowledge Management," in *PACIS 2006 Proceedings,* Kuala Lumpur, Malaysia.

Pipino, L. L., Lee, Y. W. and Wang, R. Y. 2002. "Data quality assessment," *Communications of the ACM* (45:4), 211–218.

Powers, D. M. W. 2011. *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation*.

QAS. 2013. *The Data Advantage: How Accuracy Creates Opportunity*, Experian QAS, London, UK.

Qin, X. and P. Cunningham. 2012. *Assessing the Quality of Wikipedia Pages Using Edit Longevity and Contributor Centrality*, arXiv:1206.2517.

Redman, T. C. 1997. *Data quality for the information age*, Boston, MA: Arctech House.

Redman, T. C. 2001. *Data quality: the field guide*, Boston, MA: Digital Press.

Richter, A., Stocker, A., Müller, S. and Avram, G. 2013. "Knowledge Management Goals Revisited–A Cross-Sectional Analysis of Social Software Adoption in Corporate Environments," *VINE: The journal of information and knowledge management systems* (43:2), 132–148.

Ross, S. M. 2003. "Peirce's criterion for the elimination of suspect experimental data," *Journal of engineering technology* (20:2), 38–41.

Samelson, H. 2001. "Differential Forms, the Early Days; or the Stories of Deahna's Theorem and of Volterra's Theorem," *The American Mathematical Monthly* (108:6), 522–530.

Sanders, F. 1963. "On Subjective Probability Forecasting," *Journal of Applied Meteorology* (2:2), 191–201.

Schisterman, E. F., Perkins, N. J., Liu, A. and Bondell, H. 2005. "Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples," *Epidemiology* (16:1), 73–81.

Sebastian-Coleman, L. 2015. *Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework*, Waltham, MA: Morgan Kaufmann Publishers.

Seibert, M., Preuss, S. and M. Rauer. 2011. *Enterprise Wikis:* Die erfolgreiche Einführung und Nutzung von Wikis in Unternehmen, Wiesbaden: Gabler Verlag.

Shen, A., Qi, J. and Baldwin, T. 2017. "A hybrid model for quality assessment of Wikipedia articles," in *Proceedings of the Australasian Language Technology Association Workshop 2017,* Brisbane, Australia.

Sinanc, D. and Yavanoglu, U. 2013. "A New Approach to Detecting Content Anomalies in Wikipedia," in *Proceedings of the 12th International Conference on Machine Learning and Applications (ICMLA),* Miami, FL.

Sonntag, D. 2004. "Assessing the quality of natural language text data," in *Informatik 2004–Informatik verbindet–Band 1, Beiträge der 34. Jahrestagung der Gesellschaft für Informatik eV (GI),* Ulm, Germany.

Spruit, M. and van der Linden, V. 2019. "BIDQI: The Business Impacts of Data Quality Interdependencies Model," *Technical Report Series* (UU-CS-2019:001), 1–25.

Stefanovic, D., Marjanovic, U., Delić, M., Culibrk, D. and Lalic, B. 2016. "Assessing the success of e-government systems: An employee perspective," *Information & Management* (53:6), 717–726.

Stefansky, W. 1972. "Rejecting Outliers in Factorial Designs," *Technometrics* (14:2), 469–479.

Stigler, S. M. 1980. "Stigler's Law of Eponymy," In: Gieryn, T. F. (ed.) in *Science and social structure: A festschrift for Robert K. Merton* New York, NY: New York academy of sciences, pp. 147–157.

Stocker, A. and Tochtermann, K. 2011. "Enterprise Wikis – Types of Use, Benefits and Obstacles: A Multiple-Case Study," *Communications in Computer and Information Science* (128:4), 297–309.

Strong, D. M., Lee, Y. W. and Wang, R. Y. 1997. "Data quality in context," *Communications of the ACM* (40:5), 103–110.

Stvilia, B., Twidale, M. B., Gasses, L. and Smith, L. C. 2005a. "Information quality discussions in Wikipedia," in *Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management,* Bremen, Germany.

Stvilia, B., Twidale, M. B., Smith, L. C. and Gasser, L. 2005b. "Assessing Information Quality of a Community-Based Encyclopedia," *ICIQ* (5), 442–454.

Suzuki, Y. and Yoshikawa, M. 2013. "Assessing quality score of Wikipedia article using mutual evaluation of editors and texts," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management,* San Francisco, CA.

Taylor, J. R. 1997. *An Introduction to Error analysis: The study of uncertainties in physical measurments.* Second Edition, Sausalito, CA: University Science Books.

Thode, H. C. 2002. *Testing For Normality*, Baton Rouge: CRC Press.

Thompson, W. R. 1935. "On a Criterion for the Rejection of Observations and the Distribution of the Ratio of Deviation to Sample Standard Deviation," *The Annals of Mathematical Statistics* (6:4), 214–219.

Tran, T. and Cao, T. H. 2013. "Automatic Detection of Outdated Information in Wikipedia Infoboxes," *Research in Computing Science* (70:1), 211–222.

Trkman, M. and Trkman, P. 2009. "A wiki as intranet: a critical analysis using the Delone and McLean model," *Online Information Review* (33:6), 1087–1102.

Tukey, J. W. 1970. *Exploratory Data Analysis.*, Reading, MA: Addison Wesley.

Tukey, J. W. 1977. *Exploratory Data Analysis* Vol. 33, Reading, MA: Addison Wesley.

Umbrich, J., Neumaier, S. and Polleres, A. 2015. "Quality Assessment and Evolution of Open Data Portals," in *Proceedings of the 3rd international conference on future internet of things and cloud,* Rome, Italy.

Verma, S. P. and Quiroz-Ruiz, A. 2006. "Critical values for six Dixon tests for outliers in normal samples up to sizes 100, and applications in science and engineering," *Revista mexicana de ciencias geológicas* (23:2), 133–161.

Viégas, F. B., Wattenberg, M. and Dave, K. 2004. "Studying cooperation and conflict between authors with history flow visualizations," in *Proceedings of the CHI 2004,* Vienna, Austria.

Walfish, S. 2006. "A review of statistical outlier methods," *Pharmaceutical technology* (30:11), 82–86.

Wang, R. Y., Reddy, M. P. and Kon, H. B. 1995a. "Toward quality data: An attribute-based approach," *Decision Support Systems* (13:3-4), 349–372.

Wang, R. Y., Storey, V. C. and Firth, C. P. 1995b. "A framework for analysis of data quality research," *IEEE Transactions on Knowledge and Data Engineering* (7:4), 623–640.

Warncke-Wang, M., Ayukaev, V. R., Hecht, B. and Terveen, L. G. 2015. "The Success and Failure of Quality Improvement Projects in Peer Production Communities," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing,* Vancouver, Canada.

Warncke-Wang, M., Cosley, D. and Riedl, J. 2013. "Tell me more: an actionable quality model for Wikipedia," in *Proceedings of the 9th International Symposium on Open Collaboration,* Hong Kong, China.

Wechsler, A. and Even, A. 2012. "Using a Markov-Chain model for assessing accuracy degradation and developing data maintenance policies," in *Proc. 18th Americas Conf. on Information Systems,* Seattle, WA.

Wickham, H. and L. Stryjewski. 2010. *40 years of boxplots*.

Wilcox, R. R. 2011. *Introduction to Robust Estimation and Hypothesis Testing*, San Diego, CA, USA: Elsevier Science & Technology Books.

Wilkinson, D. M. and Huberman, B. A. 2007. "Cooperation and quality in wikipedia," in *Proceedings of the 2007 international symposium on Wikis,* Montreal, Canada.

Wilrich, P.-T. 2013. "Critical values of Mandel's h and k, the Grubbs and the Cochran test statistic," *AStA Advances in Statistical Analysis* (97:1), 1–10.

Winker, P. 2010. *Empirical economic research and econometrics*, Berlin, Heidelberg: Springer Berlin Heidelberg (in German).

Wöhner, T., Köhler, S. and Peters, R. 2015. "Good Authors= Good Articles?-How Wikis Work," in *Proceedings of the Wirtschaftsinformatik 2015,* Osnabrück, Germany.

Wöhner, T. and Peters, R. 2009. "Assessing the quality of Wikipedia articles with lifecycle based metrics," in *Proceedings of the 5th International Symposium on Wikis and Open Collaboration,* Orlando, FL.

Wong, M. A. 2016. "Encylopedias," In: Bopp, R. E. and Linda C. Smith (eds.) in *Reference and Information Services:.* An Introduction,*Library and Information Science Text Series.* fourth edition, Santa Barbara, CA: Libraries Unlimited, pp. 525–554.

Xiong, M., Han, S., Lam, K.-Y. and Chen, D. 2008. "Deferrable Scheduling for Maintaining Real-Time Data Freshness: Algorithms, Analysis, and Results," *IEEE Transactions on Computers* (57:7), 952–964.

Youden, W. J. 1950. "Index for rating diagnostic tests," *Cancer* (3:1), 32–35.

Zak, Y. and Even, A. 2017. "Development and evaluation of a continuous-time Markov chain model for detecting and handling data currency declines," *Decision Support Systems* (103), 82–93.

Zeng, H., Alhossaini, M. A., Ding, L., Fikes, R. and D. L. McGuinness. 2006. *Computing Trust from Revision History*, Fort Belvoir, VA: Defense Technical Information Center.

Zhang, H., Ren, Y. and Kraut, R. E. 2020. "Mining and Predicting Temporal Patterns in the Quality Evolution of Wikipedia Articles," in *Proceedings of the 54th Hawaii International Conference on System Sciences,* Lahaina, HW.

Zhang, S., Hu, Z., Zhang, C. and Yu, K. 2018. "History-Based Article Quality Assessment on Wikipedia," in *Proceedings of the 2018 IEEE International Conference on Big Data and Smart Computing,* Shanghai, China.

Zhu, X. and Gauch, S. 2000. "Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web," in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval: ACM SIGIR,* Athens, Greece.

Zielinski, K., Nielek, R., Wierzbicki, A. and Jatowt, A. 2018. "Computing controversy: Formal model and algorithms for detecting controversy on Wikipedia and in search queries," *Information Processing & Management* (54:1), 14–36.

Zong, W., Wu, F. and Jiang, Z. 2017. "A Markov-Based Update Policy for Constantly Changing Database Systems," *IEEE Transactions on Engineering Management* (64:3), 287–300.

Ehrenwörtliche Erklärung

Ich erkläre hiermit ehrenwörtlich, dass ich die vorliegende Arbeit mit dem Titel

A probability-based approach for measuring currency in Wiki articles

selbständig angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Ich bin mir bewusst, dass eine unwahre Erklärung rechtliche Folgen haben wird.

Ulm, 26.02.2021

Lars Moestue

Lars Moestue