What Makes a Good Biography?

Multidimensional Quality Analysis Based on Wikipedia Article Feedback Data

Lucie Flekova[†‡], Oliver Ferschke[†‡], and Iryna Gurevych[†‡]

[†]Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research and Educational Information

[‡]Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt

www.ukp.tu-darmstadt.de

ABSTRACT

With more than 22 million articles, the largest collaborative knowledge resource never sleeps, experiencing several article edits every second. Over one fifth of these articles describes individual people, the majority of which are still alive. Such articles are, by their nature, prone to corruption and vandalism. Manual quality assurance by experts can barely cope with this massive amount of data. Can it be effectively replaced by feedback from the crowd? Can we provide meaningful support for quality assurance with automated text processing techniques? Which properties of the articles should then play a key role in the machine learning algorithms and why?

In this paper, we study the user-perceived quality of Wikipedia articles based on a novel Wikipedia user feedback dataset. In contrast to previous work on quality assessment which mostly relied on judgements of active Wikipedia authors, we analyze ratings of ordinary Wikipedia users along four quality dimensions (*complete*, *well written*, *trustworthy* and *objective*). We first present an empirical analysis of the novel dataset with over 36 million Wikipedia article ratings. We then select a subset of biographical articles and perform classification experiments to predict their quality ratings along each of the dimensions, exploring multiple linguistic, surface and network properties of the rated articles. Additionally, we study the classification performance and differences for the biographies of living and dead people as well as those for men and women. We demonstrate the effectiveness of our approach by the F_1 scores of 0.94, 0.89, 0.73, and 0.73 for the dimensions *complete*, *well written*, *trustworthy*, and *objective*. Based on the results, we believe that the quality assessment of big textual data can be effectively supported by current text classification and language processing tools.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Linguistic Processing*; H.3.7 [**Information Storage and Retrieval**]: Digital Libraries—*User Issues*; I.5.4 [**Pattern Recognition**]: Applications—*Text Processing*

WWW'14, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2744-2/14/04.
http://dx.doi.org/10.1145/2566486.2567972 .

1. INTRODUCTION

With the increasing importance of social aspects of the web, the Internet users of the 21st century become dialog partners rather than mere information recipients. As so-called *prosumers*, they do not only consume the content, but often participate in its production. Wikipedia, the collaboratively constructed online encyclopedia, has become the largest open-edit knowledge resource on the web, with tens of thousands daily edits. The English version alone attracts over 300 million visits per day.[1] About one fifth of its 4.3 million articles are about people. Such articles are, by their nature, prone to corruption (see e.g. John Seigenthaler [2]). Although Wikipedia defines a strict set of standards for creating a good biographical article, the expert quality control can hardly scale fast enough. Using machine learning techniques to predict article quality, we could save lots of human judgements for easier cases, while devoting human efforts only to harder cases that the system may struggle to handle. Can we effectively complement the manual process using automated text processing techniques based on user feedback? Which properties of the articles should then play a key role in the machine learning algorithms and why?

We have exploited a novel large-scale dataset of over 36 million Wikipedia article ratings, which represent the point of view of ordinary users, in contrast to previous work on quality assessment which made use of the ratings by Wikipedia contributors or the Wikipedia flaw markers, as detailed further in this paper. Wikipedia users have rated the articles in four dimensions – objectivity, trustworthiness, writing style, and completeness.

We first present an empirical analysis of our dataset. We then select a subset of articles with peoples' biographies in Wikipedia and perform classification experiments to predict their quality ratings along each of the dimensions. We further discuss the suitability of the linguistic, surface and network features used. Additionally, we perform a contrastive study of biographical articles about living and deceased people in order to analyze whether they receive different ratings by the users. We finally extend the analysis with a gender study to detect systematic differences in the ratings of articles about men and women. To our knowledge, this is the first large-scale study of its kind, providing novel insights into the crowdsourced quality assessment process and its applicability to automated quality assurance.

[1] http://stats.wikimedia.org/EN/TablesPageViewsMonthly.htm

[2] http://usatoday30.usatoday.com/news/opinion/editorials/2005-11-29-wikipedia-edit_x.htm

2. RELATED WORK

Information quality dimensions. Information quality is generally defined as "fitness of use", considering that the perception of quality is task-dependent and subjective [25]. Stvilia et al. [39] suggest a detailed information quality model tailored to fit the context of a large-scale, collaboratively created resource like Wikipedia. They define 22 quality criteria in three dimensions - intrinsic, contextual and reputational. Zhu et al. [44] assess the quality of answers on social QA sites and develop a multi-dimensional quality model consisting of 13 dimensions - Informativeness, Relevance, Usefulness, Completeness, Readability, Truthfulness, Objectivity, Conciseness, Politeness, Level of Detail, Originality, Novelty, and Expertise – of which the first three had the highest correlation with an overall quality score (0.79 - 0.92). These, however, are part of Stvilia's contextual dimension, which is dependent on the subjective information needs of the user and hence not possible to determine from our feedback data. We focus on the intrinsic quality criteria, which are the second most important after the contextual ones, with correlations to the overall quality ranging from 0.24 (readability) to 0.45 (trustworthiness) on social QA sites. These can be, in contrast to e.g. relevance, evaluated on an aggregate level. Weimer and Gurevych [41] assess the quality of forum posts extracted from `nabble.com` on three different topics. Their system combines surface, lexical, syntactic, forum specific and similarity features and outperforms the majority class baseline for all three datasets. In their case, lexical features measuring spelling error and swear word frequency had only a small impact on the overall classifier performance. They identify opinion-based quality ratings as one of the important sources of errors, i.e., the users express in their rating whether they support the author's opinion rather than if they consider the post of high quality. This makes the data noisy.

Wikipedia quality. In previous work, Wikipedia quality analysis has primarily been based on the judgments of active contributors rather than the average user. The focus was furthermore limited to either good articles or articles with specific type of quality problems. Early Wikipedia quality research started out with a small number of outstanding articles labeled as "featured"[3] or "good" [42, 40, 5, 21, 30, 24], and focuses on their distinction from the rest. The machine-based assessment performs almost perfect in most of the cases. However, featured articles differ from non-featured ones already by metrics like the number of edits [42], and an F_1-score over 0.9 can be reached even using some simple criteria such as word count [5]. This is why other researchers [2, 13] used so called "cleanup templates" - labels assigned by editors to fix various quality flaws - as examples of distinctively low quality articles rather than high quality ones. None of these experiments however answers the research question how quality is perceived by the Wikipedia users. Does their perception of quality vary from the one of active contributors? Yaari et al. [43] asked a small group of users to read five Wikipedia articles of different quality, labeled by Wikipedia editors, and to order them from best to worst based on the quality perceived. 9.38% of the users pointed the "featured" article as the worst of all five. Chesney et al. [6] conducted another user study – 54 researchers were asked to read a Wikipedia article and assess its credibility. They showed that users tend to assign lower trustworthiness to articles out of their area of expertise. Numerous studies [34, 29, 18] revealed that users with a lack of knowledge about the content topic are likely to base their trustworthiness judgement on surface features, such as the page layout.

Quality of Wikipedia biographies. In 2004, Halavais [4] intentionally entered incorrect information to 13 existing articles. His errors were all corrected within a couple of hours. Regarding biography pages, historian Roy Rosenzweig claims [35] that "Wikipedia is surprisingly accurate in reporting names, dates, and events in U.S. history. In the 25 biographies [he] read closely, [he] found clear-cut factual errors in only 4. Most were small and inconsequential." Editors of biographies must adhere to Wikipedia's core content policies, which include the *neutral point of view*, *verifiability* and *no original research*. They must provide reliable sources and respect the presumption of privacy when writing about living people. These criteria are manually reviewed by members of the *WikiProject Biography*, which manages the creation, development, and organization of biographical articles. According to these reviewers, about 60% of over one million assessed biographies "provide very little meaningful content, may be little more than a dictionary definition" and over 20% "provide some meaningful content, but most readers will need more."[5] Unfortunately, there is only a very small overlap between the articles rated in the Wikipedia *Article Feedback Tool*[6] project and in the *WikiProject Biography* in the same time frame, hence direct comparison of the ratings is not possible. Sadly, the same is true also for the Wikipedia quality flaw dataset used in the experiments of Ferschke et al [13].

Wikipedia and readability. Islam and Mehler [23] used a corpus of 645 Wikipedia articles from the Wikipedia Article Feedback ratings in category *Well written* to train a readability classifier. Using lexical, syntactic, semantic and information-theoretic features with WEKA SMO, they obtain an F-score of 0.75. In their experiments, the lexical and POS-based features outperform the syntactic and semantic ones. However, in contrast to our work, they do not account for the high correlation of the readability ratings with the other three rating categories. Hence the results can be overly optimistic due to the features capturing e.g. completeness of the article rather than the actual readability. Furthermore, the performance of each of the feature sets used increases significantly when changing from average occurrence counts per sentence to absolute counts per document, suggesting a strong length bias of the classifier.

3. ANALYSIS OF ARTICLE RATINGS

3.1 Article Feedback Tool

In September 2010, the Wikimedia Foundation introduced the *Article Feedback Tool* (AFT), a project for gathering article feedback from Wikipedia users. It allows the whole Wikipedia community to evaluate articles along the dimensions *Trustworthy*, *Objective*, *Well written* and *Complete* on a five-star scale. The user interface is displayed in figure 1. In July 2011, the AFT has been deployed to the whole English Wikipedia.

For our experiments, we use a publicly available dataset with over 36 million ratings for more than 1.5 million articles gathered between July 2011 and July 2012[7]. We furthermore employ a dataset of nearly 8 million ratings collected from March to September 2011, retrieved from the Wikimedia Toolserver[8], which contains additional information about the raters, such as the level of expertise

[3] `http://en.wikipedia.org/wiki/WP:FA`

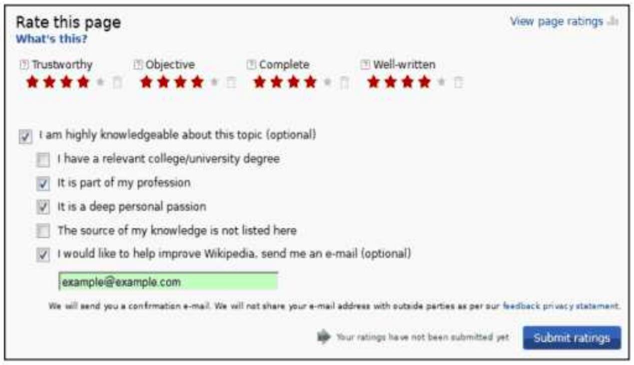
[4] `http://alex.halavais.net/the-isuzu-experiment`

[5] `http://en.wikipedia.org/wiki/WP:BIOG/A`

[6] `http://en.wikipedia.org/wiki/WP:Article_Feedback_Tool`

[7] `http://datahub.io/dataset/wikipedia-article-ratings`

[8] `http://toolserver.org/`


Rate this page
What's this?
View page ratings

Trustworthy ★★★★☆
Objective ★★★★☆
Complete ★★★★☆
Well-written ★★★★☆

☑ I am highly knowledgeable about this topic (optional)
□ I have a relevant college/university degree
☑ It is part of my profession
☑ It is a deep personal passion
□ The source of my knowledge is not listed here
☑ I would like to help improve Wikipedia, send me an e-mail (optional)

example@example.com

We will send you a confirmation e-mail. We will not share your e-mail address with outside parties as per our feedback privacy statement.

Your ratings have not been submitted yet

Figure 1: Wikipedia Article Feedback box (Version 4) as it appeared on article pages

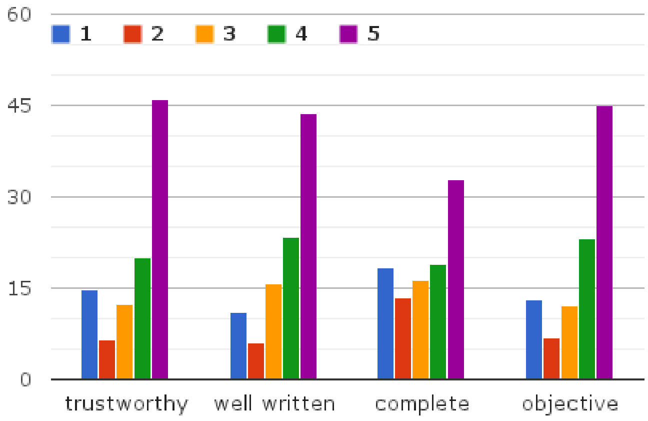


Figure 2: Percentage of rating scores (1–5) per category

in the domain and the individual user ID. After the initial dataset analysis we restricted our experiments to the biographical articles, in order to reduce the previously observed [13] bias of article style (given by the character of the topic) to our classifier.

3.2 Corpus Statistics

As displayed in figure 2, raters tend to rate articles with positive scores. They express their satisfaction (4 or 5 stars) 3–4 times more often than their dissatisfaction (1 or 2 stars). The difference is slightly lower for the dimension *Complete*, where four- and five-star ratings constitute only 52% of the ratings. However, the average ratings per article are normally distributed, with a positively shifted mean. Figure 3 illustrates this positive shift of the distribution of average article rating scores, which we observed in all four dimensions. This is in contrast with the point of view of experts from the *WikiProject Biography*, who considered the majority of reviewed biographies unsatisfactory. We explain this phenomenon partly by the fact that the experts focused on reviewing newly created, low quality articles, while the feedback from users is rather given to more popular, more often visited biographies, which are likely to be more frequently edited and improved. The pairwise correlations between all dimensions are moderately high, with a sample correlation coefficient between $r = 0.66$ (*Objective* v. *Complete*) and $r = 0.74$ (*Objective* v. *Trustworthy*) as displayed in table 1. If we consider the average rating score of all four dimensions as an overall quality score of an article, the dimensions *Trustworthy* and *Well written* show higher correlation with this score ($r = 0.89$) than the dimensions *Objective* and *Complete* ($r = 0.87$).

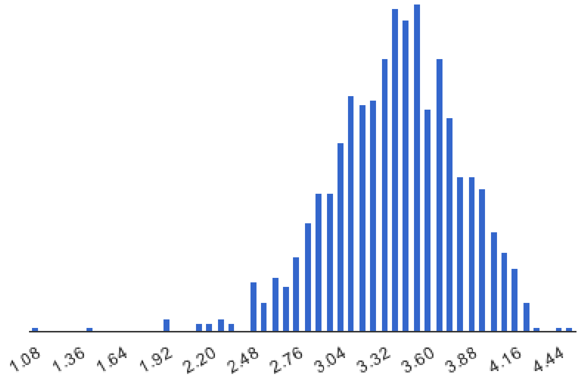


Figure 3: Histogram of average rating scores per article for dimension *"well written"*

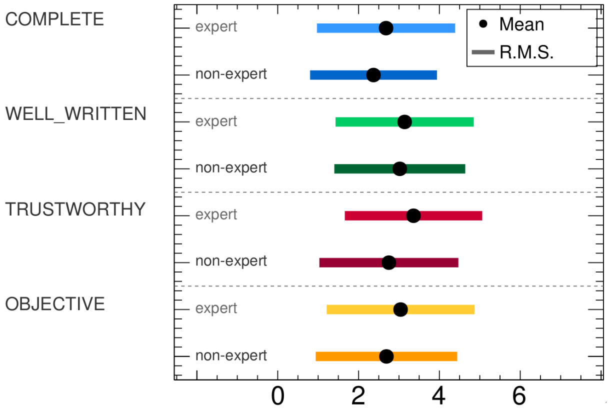


Figure 4: Expert vs non-expert ratings per dimension

The ratings per article remain consistent over time, i.e. they do not display stable increasing or decreasing trends across the revision history. The rating average for "featured" articles ranges from 3.79 (*Objective*) to 3.96 (*Well written*), only few decimal points above the Wikipedia average.

Based on the additional information which we retrieved from the Wikimedia Toolserver, 24% of all raters claim to have certain expertise in the domain of the article they evaluated. These experts are, on average, less critical in their ratings, as displayed in figure 4. This phenomenon has already been observed in the Wikipedia user study conducted by Chesney [6].

A detailed review of the the Wikimedia Toolserver data revealed that 1,834 articles (5,466 distinct article revisions) obtained five-star ratings multiple times from the same user ID. This mainly affected articles about institutions and people, such as non-accredited universities, members of political families or pop idols. However, such articles represent less than 1% of the dataset. The average rater provided feedback for 1.1 articles and 98.7% of all raters rated up to 3 articles. The largest number of articles rated by one user was 1,682, which mainly involved five-star ratings of articles about U.S. politicians.

Table 1: Pair correlation of ratings in the analyzed dimensions

	Trustworty	Objective	Complete	Well written
Trustworthy	1	-	-	-
Objective	.745	1	-	-
Complete	.702	.660	1	-
Well written	.720	.704	.737	1

Table 2: Statistics of the Wikipedia article feedback dataset

	Trust.	**Obj.**	**Comp.**	**Well w.**
No.of articles				
All topics	1,430,644	1,366,152	1,408,652	1,423,247
Living people	230,836	218,261	222,810	224,608
Dead people	128,831	121,535	125,667	128,014
Avg rating score				
All topics	3.61	3.66	3.08	3.65
Living people	3.65	3.61	3.02	3.64
Dead people	3.57	3.64	3.02	3.65
Avg no.of ratings				
All topics	6.49	6.28	6.40	6.85
Living people	5.56	5.35	5.46	5.89
Dead people	5.96	5.86	5.88	6.38

4. EXPERIMENTAL SETUP

For our classification experiments, we select a subset of the dataset with ratings for articles from the Wikipedia categories *Living people*, *Dead people* and any of their subcategories. An overview of the number of available articles per main category is given in table 2. One of our reasons for focusing on biography articles only is to reduce the impact of some article properties given by the character of a topic itself. For example articles describing fiction can be problematic when pretending to talk about reality, while for other articles this is not an issue.

Taking into account the high deviation in the rating scores per article, we have approached the problem as a binary classification task rather than distinguishing each of the five rating degrees. We define two classes of ratings, `high` and `low`, with a rating average per article of at least 4 being considered as high and an average of at most 3 as low, thus capturing the shifted distribution as illustrated on Figure 3. The corpus selection and the classification is performed separately for each of the four rating dimensions (*Complete, Objective, Trustworthy, Well written*). In each of the observed dimensions, we consider only articles with at least 10 ratings.

The number of articles fulfilling the selection criteria above is listed in table 3. We complemented this training selection with manual gender annotation. Approximately a quarter of the selected articles are biographies of women, while the remaining three quarters are biographies of men. This proportion holds true for both living and dead people. The corpus is freely available online[9].

From each of the article groups listed in table 3, we select 1,000 articles of class `high` and 1,000 articles of class `low` in each of the four observed dimensions. To account for the high correlation between dimensions, we first pick articles that appear only in one of the four dimensions of the training corpus. This means that they have outstanding average ratings only in the observed dimension, while in the other three they are average. In cases where this approach does not provide us with enough data (e.g. in `high` completeness or `low` writing style), we select the remaining articles randomly from the training corpus until we reach 1,000 articles.

Table 3: Number of documents selected for the training corpus

	Trust.	**Obj.**	**Comp.**	**Well w.**
High rating				
Living	8,453	6,863	3,809	9,878
Dead	3,635	3,015	2,033	5,117
Total	12,088	9,878	5,842	14,995
Low rating				
Living	2,571	2,700	6,145	1,836
Dead	1,514	1,346	2,727	911
Total	4,085	4,046	8,872	2,747

Our experimental setup is based on DKPro Core[10], an open-source natural language processing toolkit building upon the Unstructured Information Management Architecture (UIMA) [12]. As a runtime environment, we use the DKPro Lab [10], which allows to combine NLP pipelines into one configurable and modular system. We furthermore use the open-source API JWPL [14] to access Wikipedia based on a database dump from 1st April 2012. For preprocessing the articles, we use the components integrated in DKPro Core for sentence splitting, tokenization, stop-word annotation, readability annotation, POS tagging, lemmatization, chunking, spell check and named entity recognition, further detailed in corresponding feature descriptions. Finally, feature extraction is performed using the ClearTK framework [32]. Additional details on the system architecture can be found in the description of the FlawFinder framework [13], which is a predecessor of the system used in our experiments. As opposed to our system, it focuses mainly on structural Wikipedia features and the temporal distinction between article revisions.

We conduct a detailed feature analysis using the Weka Knowledge Explorer[11] and the ROOT data analysis framework[12].

We perform binary classification using machine learning algorithms from the Weka toolkit. The best performance with a high number of features was reached using an SVM-SMO classifier with a gaussian RBF kernel with parameters empirically set to $C = 1$ and $\gamma = 0.1$. With small feature sets (i.e., when lexical features were excluded), support vector machines were outperformed by the AdaBoostM1 algorithm [19] with decision stumps as a weak learner.

The performance of the classifier has been evaluated with stratified ten-fold cross validation on 2,000 articles from training sets of each dimension, split equally into positive and negative instances as described above. Results are compared to the majority class baseline (of the equal split) with regards to accuracy, precision, recall and F-score. For each of the dimensions, we also evaluate the performance of the classifier using each of the feature classes (lexical, syntactic etc.) separately. To assess how well our system generalizes, we additionally measure its performance with first all Wikipedia-specific features excluded, then all language-dependent features excluded.

[9] `http://www.ukp.tu-darmstadt.de/data/quality-assessment`

[10] `http://dkpro-core-asl.googlecode.com`

[11] `http://www.cs.waikato.ac.nz/~ml/weka/gui_explorer.html`

[12] `http://root.cern.ch`

5. FEATURES

We divide the features into 9 classes described below. We use the Information Gain feature selection approach [28] to rank and prune the feature space, using the top 3,000 features as the best experimental trade-off between the processing speed and the classification performance.

Surface features and readability measures. To capture the surface properties of text, we measure the length of documents, sentences and words and their proportions to each other. We measure sizes of article sections and ratios between them as well. Since previous research on readability [20, 17] has shown that shorter words are easier to read, we also count the ratio of words longer than five letters and words shorter than three letters compared to all words. We also calculate the type token ratio on lemmas created with the Gate lemmatizer [9]. The type token ratio indicates the lexical density of text and is considered to be a readability predictor. It can, however, get easily influenced by the text length, as longer texts repeat more words. We implemented the most prominent readability metricssuch as the Flesch-Kincaid Grade Level [26], the Automatic Readability Index [38], the LIX Index [4], the Coleman-Liau Index [7] and the Flesch Reading Ease [17]. The majority of those is computed using the average word and sentence lengths and number of syllables per sentence, combined with manually determined weights. The SMOG grade [31] and the Gunning-Fog Index [20] also consider the number of complex words defined as words with three or more syllables. More recent research on readability [11] shows, however, that longer sentences do not necessarily have to be more complex in terms of syntax and that other properties of the text, such as POS tags, should be taken into account.

Lexical features. We use word unigrams, bigrams and trigrams, extracted from the text before lemmatization and cleaned with the Snowball stopword list [13]. The ngrams capture not only the topic of the article, but also the tone-of-voice differences of content words, e.g."slaughtered" vs "killed". We extract n-grams not only from the article, but separately also from the associated article Talk page. We measure the information-to-noise ratio as described by Stvilia et al. [40].

Sentiment features. Inspired by the author profiling research [16], we adopted several features which were previously successfully used to indicate authors with more mature, more objective writing style. We measure the occurrence of emotionally charged adjective and adverb endings (*-ly, -ful, -ous*, ...) introduced by Corney et al. [8], based on the assumption that less objective texts may use more emotional words such as *fabulous* or *awfully*. We furthermore use word lists[14] related to anger, sadness, fear, happiness and sensations, as we expect articles written by mature authors (and therefore having good quality) to be less emotional [33]. We finally measure the proportion of words expressing certainty (e.g., *surely*,*absolutely*) and uncertainty (e.g., *perhaps*,*supposedly*) and we extract exclamation point and question mark patterns.

Spelling errors. To capture the writing style, but potentially also less trustworthy authors, we count the number of spelling errors for each part-of-speech type using the Jazzy spell checker[15] with its DKPro wrapper. We consider only the spelling error candidates which have no more than one letter different from the vocabulary in order to avoid marking names and expert terms as misspelled common nouns.

Semantic features. The named entity features capture the number of named entities in the article, using the Stanford Named Entity Recognizer [15], in particular the 3-class model with distributional similarity features for tagging all entities of the types Person, Organization and Location. We use both the overall named entity counts and the average number of named entities per sentence as features, assuming that an article with more named entities is more difficult for the reader to follow [11], but on the other hand could be perceived as more complete or more trustworthy.

Syntactic features and punctuation. We measure the ratio of each POS type as labeled by the OpenNLP POS tagger[16] and the ratios of each chunk type annotated by the TreeTagger Chunker [37]. These features are traditionally used to capture writing style [3, 36], as a text with more nouns and articles compared to pronouns and adverbs is considered more formal [27]. A large number of nouns can also make the text harder to follow [1], as the user may need to associate the co-references with more entities. We implement the contextuality measure [22], comparing implicitness and explicitness of the text based on POS tags used:

$$\frac{(nouns + adj. + prep. + art.) - (pron. + verbs + adv. + interj.)}{2 \sum POS}$$

In order to capture neutrality, we detect the ratios of comparative and superlative adjectives and adverbs. We also measure the ratio of future and past verb tenses and singular and plural nouns and pronouns, assuming, that more trustworthy information may be rather formulated in singular (as opposite to expressions like *some people say*). As a readability clue, we retrieve the proportions of inner punctuations, end punctuations and commas, with the hypothesis that longer sentences, which include more punctuation, can be harder to read. We count occurrences of numbers in the text as a possible completeness indicator.

Network features. Additional features capture the number of links to and from other articles, the presence of links from other articles, the number of links to other articles normalized per sentence, the number of outgoing links outside Wikipedia and the number of links to other language versions of the article.

Article revision history. We measure the article age, the number of article revisions, the number of unique contributors and the number of contributors registered.

References to sources. Reference-based features capture if the article has reference lists, if it includes *citation needed* tags assigned by the editors and how many references are in the articles, as well as their ratio per sentence and per text.

In addition to the individual feature groups, we experimented with three combined feature settings: *All but Wikipedia*, *All but lexical* and *All but language*. The *All but Wikipedia* setting includes all features except the last three groups (network, revision history, references) and except the wikipedia-specific surface properties such as ratio of lead paragraph to the rest of the article. The *All but lexical* group excludes both the article n-grams and the article Talk page n-grams. The *All but language* settings includes all Wikipedia-specific and surface features, but excludes all features

[13] http://snowball.tartarus.org/

[14] http://www.enchantedlearning.com

[15] http://jazzy.sourceforge.net/

[16] http://opennlp.sourceforge.net/

Table 4: Features with the highest information gain (IG) in each feature class in the run with all features

Trustworthy	IG	**Objective**	IG	**Well written**	IG	**Complete**	IG
Surface features		**Surface features**		**Surface features**		**Surface features**	
No. of discussions	.0218	Readability-Col.Liau	.0712	No. of tokens	.3320	No. of tokens	.6990
Ratio of short words	.0164	No.of characters	.0492	No. of characters	.3218	No. of characters	.6908
No. of sentences	.0143	Ratio of long words	.0452	No. of sentences	.2940	No. of sentences	.6599
Content-based feat.		**Content-based feat.**		**Content-based feat.**		**Content-based feat.**	
Ending *-ible*	.0129	Number ratio	.0523	Ending *-ous*	.1609	Sensation words	.4062
		Modal verb *should*	.0431	Ending *-able*	.1500	Ending *-able*	.3935
		Negative words	.0182	Positive words	.1487	Positive feelings	.3844
Spelling features		**Spelling features**		**Spelling features**		**Spelling features**	
Spelling error ratio	.0134			Pronoun error ratio	.1471	Card.error ratio	.3706
				Prep.error ratio	.1398	Prep.error ratio	.3628
				Adverb error ratio	.1206	Pronoun error ratio	.2940
Syntactic features		**Syntactic features**		**Syntactic features**		**Syntactic features**	
Ratio of articles	.0338	Ratio of prepositions	.0674	Superlative adv.ratio	.1580	Superlative ratio	.4309
Ratio of nouns	.0226	Verb chunk ratio	.0590	Superlative adj.ratio	.1415	Question ratio	.2765
Ratio of prepositions	.0202	Noun ratio	.0404	Present verb tense	.1208	Inner punctuation	.2755
Ratio of verbs	.0188	Superlative adj.ratio	.0397	Pronoun ratio	.1176	Present verb tense	.2734
Ratio of plurals	.0179	Inner punctuation	.0296	Inner punctuation	.0997	Exclamation rate	.2379
Ratio of superlatives	.0169	Plural ratio	.0142	Question ratio	.0997	Interjection rate	.1741
Network features		**Network features**		**Network features**		**Network features**	
No. of languages	.0195	No.of external links	.0469			No. of languages	.4238
History-based feat.		**History-based feat.**		**History-based feat.**		**History-based feat.**	
No. of reg.contrib.	.0175	Article age	.1412	Article age	.5899	Article age	.6230
				No. of reg.contrib.	.3484	No. of reg.contrib.	.5740
				No. of contributors	.3313	No. of contributors	.5440
				No. of revisions	.2894	No. of revisions	.5500
Reference-based f.		**Reference-based f.**		**Reference-based f.**		**Reference-based f.**	
No. of references	.1129						
Lexical features		**Lexical features**		**Lexical features**		**Lexical features**	
election	.0379	political	.0794	time	.1452	time	.3030
political	.0353	policy	.0725	including	.1394	made	.2907
award	.0316	government	.0690	early	.1351	left	.2890
policy	.0268	politicians	.0638	character	.1329	end	.2768
rights	.0267	leaders	.0635	world	.1307	early	.2757
voters	.0265	accused	.0604	film	.1304	including	.2651
senate	.0248	category actors	.0600	night	.1300	thumb	.2745
claims	.0243	actors	.0591	made	.1250	part	.2643
party	.0242	election	.0598	play	1235	back	.2630
elected	.0240	minister	.0551	title	.1234	years	.2619

dependent on the English language, i.e. lexical, syntactic, semantic, spelling and sentiment features.

6. EXPERIMENTAL RESULTS

The precision, recall and F_1-scores of the cross-validation with alternated feature sets in each feedback dimension are displayed in table 5. The best performance for the sets with lexical features included was reached using an SVM-SMO classifier with a gaussian RBF kernel while with small feature sets (i.e. when lexical features were excluded), the reported performance is achieved with the AdaBoostM1 algorithm using decision stumps as a weak learner. Supposedly, the AdaBoostM1 suffered from overfitting the training data when we included the lexical features. In our terminology, instance is classified positively when a quality problem is found, and the precision and recall is reported from this perspective.

The features with the highest information gain in each dimension are listed in table 4. An extended version of this table, together with our histogram visualizations of the feature values for the two classes, are the background for the analysis in this section.

The biography dataset, which we have used in our core analysis, copies the Wikipedia distribution of the age and gender of portraited people, containing approximately one third of the biographies of dead and two thirds of living people, of which about one quarter were women. We additionally performed the classification experiments for 2,000 living and 2,000 dead people's biographies separately, using all of the feature groups together. In a similar way, we also trained the quality classifiers of male and female biographies in our mixed dataset for both genders separately. Observations from these experiments are described in the *contrastive study* paragraphs.

6.1 Dimension: Well written

Classification performance. For the dimension *Well written* we obtain an F_1-score of 0.89 on all biographies. In the same dimension, Islam et al. [23] achieved an F-score of 0.75 on an unspecified subset of 645 articles from the feedback dataset, using mainly lexical and POS-based features. We reach the best results using either spelling error features only or the Wikipedia-specific network features or historical information, such as the number of incoming links or the number of article revisions. Competitive score (0.88) is achieved with sentiment-based features. With each

Table 5: Classification performance for the low quality class

Feature group	F_1-score	Prec.	Recall	Accur.
Well written				
Network-based	.888	.934	.847	.898
Spelling errors	.886	.944	.834	.897
History-based	.885	.947	.829	.896
Sentiment	.877	.928	.832	.879
All but lexical	.823	.819	.828	.829
References	.816	.939	.873	.886
All but wikipedia	.812	.859	.770	.835
Lexical	.810	.859	.767	.724
All but language	.802	.789	.816	.806
All	.792	.733	.863	.803
Syntactic	.791	.767	.816	.775
Surface & readability	.749	.636	.910	.678
Talk page n-grams	.716	.746	.689	.739
Baseline	.666	.500	1.00	.500
Trustworthy				
History-based	.727	.665	.802	.684
References	.723	.678	.775	.639
Network-based	.714	.660	.718	.639
Sentiment	.699	.675	.724	.674
Lexical	.668	.663	.674	.653
All	.664	.661	.667	.705
Baseline	.666	.500	1.00	.500
All but language	.630	.605	.657	.603
Syntactic	.621	.587	.682	.576
All but lexical	.590	.625	.558	.602
Spelling errors	.573	.588	.559	.573
Surface & readability	.555	.576	.536	.559
Objective				
All but lexical	.732	.760	.706	.701
All but Wikipedia	.713	.686	.744	.666
Lexical	.712	.682	.745	.717
Sentiment	.709	.673	.748	.744
All	.704	.680	.724	.701
Network-based	.701	.760	.650	.739
Baseline	.666	.500	1.00	.500
Surface & readability	.648	.660	.633	.677
History-based	.615	.757	.518	.695
Syntactic	.614	.621	.606	.594
Talk page n-grams	.609	.742	.516	.688
All but language	.287	.628	.186	.524
Complete				
History-based	.939	.921	.943	.927
Network-based	.937	.935	.939	.925
Sentiment	.934	.923	.939	.922
Lexical	.930	.948	.912	.919
All	.917	.892	.943	.908
Syntactic	.843	.936	.766	.845
Surface & readability	.761	.948	.636	.774
References	.680	.701	.661	.674
Baseline	.666	.500	1.00	.500

of these feature classes, we could reach a precision of over 0.93, while the highest recall (0.91) was obtained with the structural features.

Helpful features. As expected, well written articles are older, with a higher number of revisions and more contributors. They also tend to be longer, which is reflected in a lower type-token ratio. Surprisingly, the articles rated as well written are more emotionally intense, which is captured by the excess of words expressing positive emotions and feelings, superlatives, adjectives with emotional endings (e.g. *remarkable*, *memorable*) and exclamations. Highly rated articles also use more modal verbs, uncertainty expressions (*perhaps*, *possibly*, *maybe*, ...) and questions, as well as clarification expressions such as *clarify*, *specify* or *explain*, and they contain more inner punctuation.

Badly written articles contain more errors in prepositions and punctuation and verbs referring to present or future rather than to the past. They use more words expressing certainty, such as *obviously*, *clearly* or *absolutely*.

The readability features, which mainly capture the length of words and sentences, did not improve the prediction performance in this dimension. This implies that writing style ratings are based on syntactic, stylistic and semantic criteria rather than surface properties.

Contrastive study of age and gender. Well written articles about living people appear in more languages and have a smaller proportion of lead paragraph to the rest of the article. Articles with low ratings are more often short stubs. The ratings of biographies of living people are negatively correlated to the usage of the present verb tense.

Judging from the values of the n-gram features with a high information gain in both experiments, the biographies of women were more likely to have lower writing style ratings if they did not contain any words related to their personal life, such as *mother*, *father*, *married* or *love*. In contrast, low rated biographies of men often lacked the word *career*. High rated biographies of women also contained more named entities referring to persons.

6.2 Dimension: Trustworthy

Classification performance. Our classifier performs best (F_1 = 0.73) with history-based features only. Trustworthy articles, as expected, have more references, and often exist in more language versions, thus the high performance (F_1 = 0.71) of network-based features. Using lexical features (F_1 = 0.67), we observed certain topic bias in this dimension, e.g. the word ngrams such as *category indian*, *india*, *indian* appeared to be rather strong predictors for a high trustworthiness and *government* or *politicians* for a low one.

Helpful features. The trustworthiness of an article improves with its age and number of contributors. Trustworthy articles, as expected, have more references, and often exist in more language versions. Trustworthy articles often contain more nouns, while pronouns, verbs and definite and indefinite articles would appear more often in the biographies with lower trustworthiness ratings. This information alone, however, is not sufficient as it yields only an F_1-score of 0.62.

Articles with low trustworthiness have a higher ratio of emotionally intense adjectives in the text, and use more words expressing certainty. They contain less modal verbs and inner punctuation. These articles also usually have more comments on the article talk page, demonstrating the efficiency of the community-based recovery mechanism for this type of quality problem.

Contrastive study of age and gender. Biographies of living people were rated more trustworthy when more references per sentence were included and less trustworthy when having higher ratio of present to past verbs.

Low rated biographies of men contain more negative words, while low rated articles about women contained more positive emotions and third person pronouns. The biographies of men with low ratings are, judging from the n-gram features, more often related to certain conflicts (*middle east*, *holocaust*, ...), while those of women contain in most cases some overstated achievements of celebrities (*award*, *greatest*, *star*, *successful*, ...).

6.3 Dimension: Objective

Classification performance. Objectiveness is the only dimension in which the classification result of our system is better when we ignore all Wikipedia-based features (such as the number of links, references, revisions or article age). We obtain the highest F_1-score (0.73) by combining stylistic, syntactic, surface (length-based) and sentiment-based features. The Wikipedia-based features contribute to a high precision of the classifier, but result in a low recall.

Helpful features. Articles with low objectivity ratings had a higher ratio of long words and thus a lower readability score (Coleman-Liau measure). These articles also show higher proportion of first and third person plurals, prepositions and verbs. They contain more often the word *should*, more negative words and terms expressing anger. However, they also contain more expressions associated with positive emotions and feelings. Furthermore, named entities referring to organizations appear more often in such articles. Finally, objective articles tend to be older, contain more nouns and, surprisingly, more superlatives.

The most predictive word n-grams in this dimension could be grouped into three distinct categories – politicians, actors and musicians. While words related to politics are good predictors of problematic articles, expressions associated with the film and music industry are predictive for articles rated as highly objective. Low objectivity occurs more often for people in the Wikipedia categories of the type “Members of ... ” (party), which is most likely related to the politician bias.

Contrastive study of age and gender. In comparison to deceased people, low rated biographies of living people are more correlated with n-grams such as *politicians* and *elections* and high rated biographies tend to have more language versions. In the biographies of dead people, positive words and expressions of sadness were predicting an objective article, which was not the case for the articles about living people. This makes us wonder if the users value the principle of the saying “Of the dead, nothing unless good” - however, we cannot disprove that this observation might be only corpus-specific.

The low rated articles about men contain more long words and words referring to politicians (both republicans and conservatives, as illustrated in figure 5), words referring to government or high amounts of money (*million*, *billion*). In the case of biographies of women, the word *hot* indicates highly rated articles with a surprisingly high information gain (0.08). A detailed manual analysis revealed, that objective articles about women include many biographies of actresses including a broad coverage of the history of pornography, which seemed to be the main source for these kinds of predictive lexical cues.

6.4 Dimension: Complete

Classification performance. As table 5 shows, we reach an F_1-score of 0.94 already with a small number of simple features (article age, number of revisions, number of contributors). Similar results can be obtained with almost all other individual feature groups except of syntactic, surface and reference features.

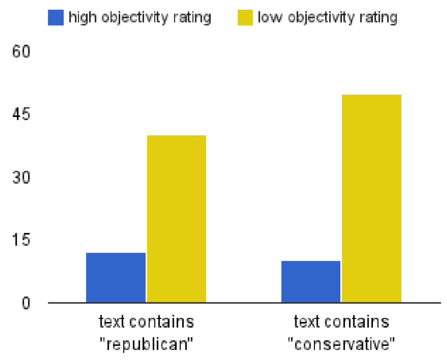


Figure 5: Number of high rated (blue color) and low rated (yellow color) articles about men, containing the words *republican* and *conservative*

Helpful features. Ratings in this dimension can be well predicted without linguistic features. High performing results obtained with the n-gram based or network-based features or sentiment analysis are in our interpretation just capturing the fact, that articles with non-zero feature values of all of the observed phenomena (e.g. containing at least one noun, one superlative, at least one negative word, etc.) are perceived as more complete. Features based on the document surface (e.g. length), text syntax, or number of references have high precision of the results, but low recall (i.e. many long documents, which are however rated as incomplete, would not be retrieved). Higher recall is achieved using the history-based features, as the complete articles are in most cases older.

Contrastive study of age and gender. In this dimension, we found no remarkable differences between the distinct biography groups.

7. ERROR ANALYSIS

7.1 Sources of false negatives

The following two sections present the common sources of errors in our dataset based on the manual analysis of misclassified instances. In our terminology, instance is classified positively when a quality problem is found. Hence when we discuss the sources of false negatives further in this section, we are talking about low rated (problem-containing) articles that were classified as high quality (problem-free) ones. If a specific dimension is not discussed explicitly, the error source was common for all four dimensions.

Controversial topics. During the manual inspection of the misclassified articles which were supposed to represent the class `low` (i.e. having the average rating of 3 or less), we found out that many of them have very bipolar ratings, oscillating between one and five stars with very low user agreement (on false negatives, Krippendorff’s $\alpha \leq 0.1$). These articles mainly discuss controversial people, e.g. fanatic religious visionaries, political extremists or pseudoscientists. Based on the talk pages, people could not reach a consensus on what is a neutral point of view on the topic, which projected not only into the objectiveness ratings but also to the other three dimensions.

Missing information. Based on the article texts and the discussion on its Talk pages, we believe that many articles were rated low in the dimension *Well written* not due to its writing style itself, but because the users were looking for some specific facts about the person, which were not contained in the article. In theory, this problem shall be captured only by the dimension *Complete*. Unfortunately, the definition of the dimensions is not clear-cut and is sometimes misinterpreted by the users. Hence the rating usually propagated to the dimension *Well written* as well.

Short-term vandalism. Since the user feedback was collected during a one-year period, the articles were undergoing frequent edits. As we have not observed a clear shift in rated quality of the individual articles between the beginning and the end of this period and as the ratings were too sparse to obtain a reliably annotated dataset using only ratings of a specific article revision, we have used only one Wikipedia snapshot. However, some articles suffered from vandalism (e.g., 8 June 2012: *"Viktor Brack, an odious bastard never should be born..."*). Although such edits were usually corrected within the same day, they occasionally projected into a sudden drop of ratings from five stars to one, decreasing the rating average of the article and thus leading to assigning a misleading gold standard label.

Repetitive information. According to the Talk pages, some of the articles presumably received low ratings in the dimension *Well written* for being written as high-school essays, containing a lot of repetitive information. This problem could not be captured correctly by the type-token ratio feature due to a largely varying length of the articles. The classifier could possibly benefit from introducing some sentence similarity measures to capture repetitive sentences.

Confusion of dimensions. The error analysis on instance level confirmed what we could already suspect from the correlation matrix (see table 1). Users struggled to distinguish easily between rated dimensions. Many biographies, which were lacking some important information about a person's life, were considered as incomplete, therefore biased, therefore badly written and not trustworthy. Articles that are badly written, but at the same time objective, trustworthy and complete are rare to find (about 6% of all available data). The same holds for complete, but at the same time biased, untrustworthy and badly written ones (about 2% of all available data). Although the cases with distinct ratings in one dimension were overrepresented in our corpus (20%) to better train for the individual dimensions, the performance of each of the four independent classifiers possibly still suffered from not being able to capture these implicit relations. Building a meta classifier, which incorporates the implication rules between dimensions, could lead to a better prediction of the ratings as well as of the overall article quality.

Topic bias. Using lexical features in the categories *Trustworthy* and *Objective* leads to a topic bias of the classifier, which then tends to predict all the articles related to government, elections or military as bad, while low quality biographies of the entertainment sector would be still classified as rated high. This problem can be avoided by completely excluding the n-gram features without loss of performance.

7.2 Sources of false positives

Article stubs. Short articles are a very common source of errors, not only for their length, but also because they often consist mainly of the list of works of the author. It is frequent that such a list contains many emotionally intense words, e.g. artwork titles, which leads to errors in the dimension *Objective*.

Large proportion of foreign language. Several misclassified biographies were very dense in the named entity content and description of places, customs or movie names in their original, non-English language. This causes errors in the dimension *Well written*, as the POS tagger and sentence parser fail, and also the ratio of potential spelling errors is artificially high. This problem could be addressed by a language recognition preprocessing step as well as more advanced treatment of named entities, which shall be excluded from other features.

Ancient people. It turned out that one of the Wikipedia subcategories of dead people biographies is `Ancient people`, which contains also, for example, biblical individuals. Entries about such persons, although not very frequent in our dataset, differ in the language style they are written in. These biographies can tolerate more poetic language and examples from ancient texts, which is possibly misleading for our classifier.

7.3 Impact of Length

The length bias of high quality articles, which was often observed in previous research [5, 21], remains an open problem. The high performance of our classifiers on the *Completeness* dimension suggests that even the features normalized per word count tend to capture this property. Similarly to the problem with the type-token ratio decreasing for longer texts, the likelihood that some of the monitored words (e.g., in the sentiment analysis or spelling correction) will be seen at least once in the text is higher when the text is longer. We are still able to correctly and meaningfully classify the articles with a high density of sentiment words. However the distinction between 0.00% and 0.01% of positive words contained in the text can signal a mere text length difference. The same applies for the measures normalized per sentence, as the likelihood that at least one sentence in the long text will have a non-zero value is higher. This comparison issue cannot be avoided by cutting the long articles to the length of the short ones either, as this is likely to cause a stylistic bias - the introduction into one topic would be compared to the core body of another. A construction of a Wikipedia quality corpus of articles with naturally equal length would address this issue - unfortunately, this could not be reliably achieved with our rating data.

8. CONCLUSIONS AND FUTURE WORK

We have analysed the four dimensions of quality – writing style, objectiveness, trustworthiness and completeness – from the perspective of users in a corpus of Wikipedia biographies. Our features capture the linguistic dimensions of the text – lexical, syntactic and semantic – as well as the surface properties of the article with and without the Wikipedia-related properties. In all the dimensions except objectiveness, the age of an article was a strong predictor of high ratings, confirming the efficiency of the collaborative quality improvement process. Articles with high ratings are also better connected with the rest of Wikipedia. We confirmed the influence of the number of edits on the article quality, as previously observed on a different dataset [42]. The length bias of high quality articles [5, 21] mainly contributed to the recall in the writing style dimension and the precision in the completeness dimension. In contrast to Ferschke et al. [13], whose system achieves the best performance in predicting neutrality and style flaws based on the Wikipedia quality templates using mainly n-gram features, we were able to achieve the best F_1 score in all four dimensions by excluding the n-gram features from our experimental setup and applying the AdaBoostM1 algorithm [19] with decision stumps as a weak learner. Similar finding – a low impact of lexical features on the classifier – have been observed in the experiments by Weimer et al. [41]. In compliance with this work, we also identified opinion-biased, noisy quality ratings as one of our main sources of errors, especially on controversial topics. Our inspection of low rated article revisions corresponds with the finding of Halavais [17] that malicious edits are normally corrected within hours from their creation.

We explored the differences between the observed quality of biographies of living and dead people and discovered that a biography of a living person is rated higher when it exhibits more past tense than present tense sentences. That could be an indicator of Wikipedia notability criteria - the biographies of people with only a few past achievements are possibly rated lower. Furthermore, high quality articles of living people usually exist in more language versions, while for dead people this is not necessarily the case. In our interpretation there might be a higher consensus on the notability of local people from the past, while the descriptions of achievements by contemporary local celebrities can be perceived as exaggerated boasting.

Articles rated as well written are, beside being older and longer, more emotionally written, express more uncertainty and are less likely to contain an excessive number of spelling errors. While we reach the highest precision using spelling-, sentiment- and Wikipedia-based features, the best recall is reached with surface features only. We achieve a good classification performance of $F_1 = 0.81$ even when all the Wikipedia-specific properties are excluded, indicating the generalizability potential of our writing style quality prediction system to other online texts such as blogs or news portals.

Trustworthy articles can be well predicted ($F_1 = 0.72$) using the information about the number of references included. Furthermore, we observed that the articles with low trustworthiness use more certainty expressions, more emotions and more pronouns rather than nouns, which is possibly perceived as less explicit by the readers. The F_1-score of 0.67 can be also achieved with lexical features only, which mainly indicate the article topic.

Objective biographies usually describe people from the cultural scene (actors, singers) while any relation to politics is a clear lexical indicator for a low objectivity rating. Low rated articles are also less readable, contain more long words, more verb chunks, certainty expressions and emotions. Biographies of politicians with low ratings are in most cases about men.

The completeness ratings could be predicted with very high precision (0.921) and recall (0.943) using only history-based features. However, the most of our experimental results in this dimension are not easily interpretable. We believe that the aspects of the completeness ratings, which cannot be correctly captured by simple descriptive features, are contextual for the user in terms of information quality and thus difficult for us to account for.

In our experiments, we were able to predict different quality aspects of Wikipedia biography articles using a large variety of features, ranging from fast and simple heuristics with Wikipedia-specific properties to largely generalizable methods analyzing the writing style. Based on the results, we believe that the quality assessment of big textual data can be effectively supported by current text classification and language processing tools.

In the future, we intend to synthesize our system into one robust quality classifier and possibly integrate it more closely with Wikipedia in order to identify candidate articles for quality improvement. While we can predict the likelihood of a trustworthiness problem based on the Wikipedia-specific properties and sentiment levels in the article, we probably would not be able to recommend trustworthiness corrections on a fine-grained text level (i.e., a specific statement missing reference), neither evaluate missing information in an incomplete article. Making this kind of predictions would require a very large amount of real world knowledge, which is hardly possible to obtain without the encyclopedia itself. Our system could, however, make valuable edit suggestions to improve objectivity or writing style of an article through replacing emotional expressions, making spelling corrections or marking content with low readability or high repetitiveness. Such a system would then be easily generalizable on any informative text outside Wikipedia. A drawback to any future work is the discontinuation of the Wikipedia Article Feedback project, since we cannot obtain additional training and testing data from the same resource.

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship program under grant no. I/82806, by the German Research Foundation under grant no. GU 798/14/1, and by the Hessian research excellence program "Landes-Offensive zur Entwicklung Wissenschaftlich-Ökonomischer Exzellenz" as part of the research center "Digital Humanities".

9. REFERENCES

[1] S. Aluisio, L. Specia, C. Gasperin, and C. Scarton. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9, 2010.

[2] M. Anderka, B. Stein, and N. Lipka. Predicting quality flaws in user-generated content. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '12*, page 981, Aug. 2012.

[3] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123, Feb. 2009.

[4] C. Björnsson. *Läsbarhet: Lesbarkeit durch Lix. (Aus dem Schwedischen).* (Pedagogiskt Utvecklingsarbete vid Stockholms Skolor. 6.). 1968.

[5] J. E. Blumenstock. Size matters. In *Proceeding of the 17th International Conference on World Wide Web*, page 1095, Apr. 2008.

[6] T. Chesney. An empirical examination of Wikipedia's credibility. *First Monday*, 11(11), 2006.

[7] M. Coleman and T. Liau. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283, 1975.

[8] M. Corney, O. de Vel, A. Anderson, and G. Mohay. Gender-preferential text mining of e-mail discourse. In *Proceedings of the 18th Annual Computer Security Applications Conference*, pages 282–289, 2002.

[9] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li, and W. Peters. *Text Processing with GATE (Version 6)*. University of Sheffield, Department of Computer Science, 2011.

[17] `http://alex.halavais.net/the-isuzu-experiment`

[10] R. Eckart de Castilho and I. Gurevych. A lightweight framework for reproducible parameter sweeping in information retrieval. In M. Agosti, N. Ferro, and C. Thanos, editors, *Proceedings of the 2011 Workshop on Data Infrastructures for Supporting Information Retrieval Evaluation*, pages 7–10, Oct. 2011.

[11] L. Feng, N. Elhadad, and M. Huenerfauth. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 229–237, 2009.

[12] D. Ferrucci and A. Lally. UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. *Natural Language Engineering*, 10(3-4):327–348, 2004.

[13] O. Ferschke, I. Gurevych, and M. Rittberger. FlawFinder: A Modular System for Predicting Quality Flaws in Wikipedia. In *CLEF 2012 Labs and Workshop, Notebook Papers*, 2012.

[14] O. Ferschke, T. Zesch, and I. Gurevych. Wikipedia revision toolkit: Efficiently accessing wikipedia's edit history. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 97–102, June 2011.

[15] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370, 2005.

[16] L. Flekova and I. Gurevych. Can we hide in the web? large scale simultaneous age and gender author profiling in social media. In *CLEF 2012 Labs and Workshop, Notebook Papers*, Sept. 2013.

[17] R. Flesch. A new readability yardstick. *The Journal of applied psychology*, 32(3):221, 1948.

[18] B. J. Fogg, P. Swani, M. Treinen, J. Marshall, O. Laraki, A. Osipovich, C. Varma, N. Fang, J. Paul, A. Rangnekar, and J. Shon. What makes Web sites credible? In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 61–68, Mar. 2001.

[19] Y. Freund and R. E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, pages 23–37. Springer, 1995.

[20] R. Gunning. The fog index after twenty years. *Journal of Business Communication*, 6(2):3–13, 1969.

[21] D. Hasan Dalip, M. André Gonçalves, M. Cristo, and P. Calado. Automatic quality assessment of content created collaboratively by web communities. In *Proceedings of the 2009 joint international conference on Digital libraries*, page 295, June 2009.

[22] F. Heylighen and J.-M. Dewaele. Variation in the contextuality of language: An empirical measure. *Foundations of Science*, 7(3):293–340, 2002.

[23] Z. Islam and A. Mehler. Automatic readability classification of crowd-sourced data based on linguistic and information-theoretic features. *Computación y Sistemas*, 17(2):113–123, 2013.

[24] S. Javanmardi and C. Lopes. Statistical measure of quality in Wikipedia. In *Proceedings of the First Workshop on Social Media Analytics - SOMA '10*, pages 132–138, July 2010.

[25] J. Juran and B. Godfrey. *The quality control process*. 1999.

[26] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, DTIC, 1975.

[27] M. Koppel, S. Argamon, and A. R. Shimoni. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412, 2002.

[28] S. Kullback. *Information theory and statistics*. 1997.

[29] G. Lindgaard, C. Dudek, D. Sen, L. Sumegi, and P. Noonan. An exploration of relations between visual appeal, trustworthiness and perceived usability of homepages. *ACM Transactions on Computer-Human Interaction*, 18(1):1–30, Apr. 2011.

[30] N. Lipka and B. Stein. Identifying featured articles in Wikipedia. In *Proceedings of the 19th International Conference on World Wide Web*, page 1147, Apr. 2010.

[31] G. H. McLaughlin. SMOG grading: A new readability formula. *Journal of reading*, 12(8):639–646, 1969.

[32] P. V. Ogren, P. G. Wetzler, and S. Bethard. ClearTK: A UIMA toolkit for statistical natural language processing. In *Proceedings of the UIMA for NLP workshop at the Language Resources and Evaluation Conference*, pages 32–38, 2008.

[33] J. Pennebaker, M. Mehl, and K. Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577, 2003.

[34] R. E. Petty. The Elaboration Likelihood Model of Persuasion. *Advances in Experimental Social Psychology*, 19:123–205, 1986.

[35] R. Rosenzweig. Can history be open source? Wikipedia and the future of the past. *The Journal of American History*, 93(1):117–146, 2006.

[36] J. Schler, M. Koppel, S. Argamon, and J. Pennebaker. Effects of age and gender on blogging. In *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, pages 199–205, 2006.

[37] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, 1994.

[38] R. Senter and E. Smith. *Automated Readability Index*. AMRL-TR-66-220. 1967.

[39] B. Stvilia, L. Gasser, M. B. Twidale, and L. a. C. Smith. A framework for information quality assessment. *Journal of the American Society for Information Science and Technology*, 58(12):1720–1733, 2007.

[40] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser. Information quality work organization in Wikipedia. *Journal of the American Society for Information Science and Technology*, 59(6):983–1001, Apr. 2008.

[41] M. Weimer and I. Gurevych. Predicting the perceived quality of web forum posts. In *Proceedings of the Conference on Recent Advances in Natural Language Processing*, Jan. 2007.

[42] D. M. Wilkinson and B. A. Huberman. Assessing the value of cooperation in wikipedia. *First Monday*, 12(4), 2007.

[43] E. Yaari, S. Baruchson-Arbib, and J. Bar-Ilan. Information quality assessment of community generated content: A user study of Wikipedia. *Journal of Information Science*, 37(5):487–498, Aug. 2011.

[44] Z. Zhu, D. Bernhard, and I. Gurevych. A multi-dimensional model for assessing the quality of answers in social Q&A sites. In *Proceedings of 14th International Conference on Information Quality*, pages 264–265, Nov. 2009.