The Impact of Topic Bias on Quality Flaw Prediction in Wikipedia

Oliver Ferschke[†], Iryna Gurevych[†‡] and Marc Rittberger[‡]

† Ubiquitous Knowledge Processing Lab
Department of Computer Science, Technische Universität Darmstadt

‡ Information Center for Education
German Institute for Educational Research and Educational Information

http://www.ukp.tu-darmstadt.de

Abstract

With the increasing amount of user generated reference texts in the web, automatic quality assessment has become a key challenge. However, only a small amount of annotated data is available for training quality assessment systems. Wikipedia contains a large amount of texts annotated with cleanup templates which identify quality flaws. We show that the distribution of these labels is topically biased, since they cannot be applied freely to any arbitrary article. We argue that it is necessary to consider the topical restrictions of each label in order to avoid a sampling bias that results in a skewed classifier and overly optimistic evaluation results. We factor out the topic bias by extracting reliable training instances from the revision history which have a topic distribution similar to the labeled articles. This approach better reflects the situation a classifier would face in a real-life application.

1 Introduction

User generated content is the main driving force of the increasingly social web. Blogs, wikis and forums make up a large amount of the daily information consumed by web users. The main properties of user generated content are a low publication threshold and little or no editorial control, which leads to a high variance in quality. In order to navigate through large repositories of information efficiently and safely, users need a way to quickly assess the quality of the content. Automatic quality assessment has therefore become a key application in today's information society. However, there is a lack of training data annotated with fine-grained quality information.

Wikipedia, the largest encyclopedia on the web, contains so-called cleanup templates, which constitute a sophisticated system of user generated labels that mark quality problems in articles. Recently, these cleanup templates have been used for automatically identifying articles with particular quality flaws in order to support Wikipedia's quality assurance process in Wikipedia. In a shared task (Anderka and Stein, 2012b), several systems have shown that it is possible to identify the ten most frequent quality flaws with high recall and fair precision.

However, quality flaw detection based on cleanup template recognition suffers from a topic bias that is well known from other text classification applications such as authorship attribution or genre identification. We discovered that cleanup templates have implicit topical restrictions, i.e. they cannot be applied to any arbitrary article. As a consequence, corpora of flawed articles based on these templates are biased towards particular topics. We argue that it is therefore not sufficient for evaluating a quality flaw prediction systems to measure how well they can separate (topically restricted) flawed articles from a set of random outliers. It is rather necessary to determine reliable negative instances with a similar topic distribution as the set of positive instances in order to factor out the sampling bias. Related studies (Brooke and Hirst, 2011) have proven that topic bias is a confounding factor that results in misleading cross-validated performance while allowing only near chance performance in practical applications.

We present an approach for factoring out the bias from quality flaw corpora by mining reliable negative instances for each flaw from the article revision history. Furthermore, we employ the article revision history to extract reliable positive training instances by using the version of each article at the time it has first been identified as flawed. This way, we avoid including articles with outdated cleanup templates, a frequent phe-

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 721–730,
Sofia, Bulgaria, August 4-9 2013.

nomenon that can occur when a template is not removed after fixing a problem in an article. In our experiments, we focus on neutrality and style flaws, since they are of particular high importance within the Wikipedia community (Stvilia et al., 2008; Ferschke et al., 2012a) and are recognized beyond Wikipedia in applications such as uncertainty recognition (Szarvas et al., 2012) and hedge detection (Farkas et al., 2010).

2 Related Work

Topic bias is a known problem in text classification. Mikros and Argiri (2007) investigate the topic influence in authorship attribution. They found that even simple stylometric features, such as sentence and token length, readability measures or word length distributions show considerable correlations with topic. They argue that many features that were largely considered to be topic neutral are in fact topic-dependent variables. Consequently, results obtained on multitopic corpora are prone to be biased by the correlation of authors with specific topics. Therefore, several authors introduce topic-controlled corpora for applications such as author identification (Koppel and Schler, 2003; Luyckx and Daelemans, 2005) or genre detection (Finn and Kushmerick, 2006).

Brooke and Hirst (2011) measure the topic bias in the *International Corpus of Learner English* and found that it causes a substantial skew in classifiers for native language detection. In accordance with Mikros et al., the authors found that even non-lexicalized meta features, such as vocabulary size or length statistics, depend on topics and cause cross-validated performance evaluations to be unrealistically high. In a practical setting, these biased classifiers hardly exceed chance performance.

As already noted above, a similar kind of topic bias negatively influences quality flaw detection in Wikipedia. Anderka et al. (2012) automatically identify quality flaws by predicting the cleanup templates in unseen articles with a one-class classification approach. Based on this work, a competition on quality flaw prediction has been established (Anderka and Stein, 2012b). The winning team of the inaugural edition of the task was able to detect the ten most common quality flaws with an average F_1-Score of 0.81 using a PU learning approach (Ferretti et al., 2012). With a binary classification approach, Ferschke et al. (2012b) achieved an average F_1-Score of 0.80, while reaching a higher precision than the winning team.

A closer examination of the aforementioned quality flaw detection systems reveals a systematic sampling bias in the training data, which leads to an overly optimistic performance evaluation and classifiers that are biased towards particular article topics. Our approach factors out the topic bias from the training data by mining topically controlled training instances from the Wikipedia revision history. The results show that flaw detection is a much harder problem in a real-life scenario.

3 Quality Flaws and Flaw Recognition in Wikipedia

Quality standards in Wikipedia are mainly defined by the *featured article criteria*[1] and the *Wikipedia Manual of Style*[2]. These policies define the characteristics excellent articles have to exhibit. Other sets of quality criteria are adaptations or relaxations of these standards, such as the *good article criteria* or the quality grading schemes of individual interest groups in Wikipedia.

In this work, we focus on quality flaws regarding neutrality and style problems. We chose these categories due to their high importance within the Wikipedia community (Stvilia et al., 2008; Ferschke et al., 2012a) and due to their relevance to content outside of Wikipedia, such as blogs or online news articles. According to the Wikipedia policies[3], an article has to be written from a neutral point of view. Thus, authors must avoid stating opinions and seriously contested assertions as facts, avoid presenting uncontested factual assertions as mere opinions, prefer nonjudgmental language and indicate the relative prominence of opposing views. Furthermore, authors have to adhere to the stylistic guidelines defined in the Manual of Style. While this subsumes a broad range of issues such as formatting and article structure, we focus on the style of writing and disregard mere structural properties.

Any articles that violate these criteria can be marked with cleanup templates[4] to indicate their need for improvement. These templates can thus be regarded as proxies for quality flaws in Wikipedia.

[1] http://en.wikipedia.org/wiki/WP:FACR
[2] http://en.wikipedia.org/wiki/WP:STYLE
[3] http://en.wikipedia.org/wiki/WP:NPOV
[4] http://en.wikipedia.org/wiki/WP:TM#Cleanup

	Flaw	Description	Articles	Templates
Neutrality	**Advert**	The article appears to be written like an advertisement and is thus not neutral	7,332	2
	POV	The neutrality of this article is disputed	5,086	10
	Globalize	The article may not represent a worldwide view of the subject	1,609	1
	Peacock	The article may contain wording that merely promotes the subject without imparting verifiable information	1,195	1
	Weasel	The article contains vague phrasing that often accompanies biased or unverifiable information	704	4
Style	**Tone**	The tone of the article is not encyclopedic according to the Wikipedia Manual of Style	4,563	6
	In-universe	The article describes a work or element of fiction in a primarily in-universe style[a]	2,227	1
	Copy-edit	The article requires copy editing for grammar, style, cohesion, tone, or spelling	1,954	6
	Trivia	Contains lists of miscellaneous information	1,282	2
	Essay-like	The article is written like a personal reflection or essay	1,244	1
	Confusing	The article may be confusing or unclear to readers	1,084	1
	Technical	The article may be too technical for most readers to understand	690	2

[a] According to the Wikipedia Manual of Style, an in-universe perspective describes the article subject matter from the perspective of characters within a fictional universe as if it were real.

Table 1: Neutrality and style flaw corpora used in this work

Template Clusters Since several cleanup templates might represent different manifestations of the same quality flaw, there is a *1 to n* relationship between quality flaws and cleanup templates. For instance, the templates `pov-check`[5], `pov`[6] and `npov language`[7] can all be mapped to the same flaw concerning the neutral point of view of an article. This aggregation of cleanup templates into flaw-clusters is a subjective task. It is not always clear whether a particular template refers to an existing flaw or should be regarded as a separate class. Too many clusters will cause definition overlaps (i.e. similar cleanup templates are assigned to different clusters), while too few clusters will result in unclear flaw definitions, since each flaw receives a wide range of possible manifestations.

Template Scope Another important aspect to be considered is the difference in the scope which cleanup templates can have. *Inline-templates* are placed directly in the text and refer to the sentence or paragraph they are placed in. Templates with a *section* parameter, refer to the section they are placed in. The majority of templates, however, refer to a whole page. The consideration of the template scope is of particular importance for quality flaw recognition problems. For example, the presence of a cleanup template which marks a single section as *not notable* does not entail that the whole article is not notable.

[5]The article has been nominated for a neutrality check
[6]The neutrality of the article is disputed
[7]The article contains a non-neutral style of writing

Topical Restriction A final aspect that has not been taken into account by related work is that many cleanup templates have restrictions concerning the pages they may be applied to. A hard restriction is the page type (or namespace) a template might be used in. For example, some templates can only be used in articles while others can only be applied to discussion pages. This is usually enforced by maintenance scripts running on the Wikimedia servers. A soft restriction, on the other hand, are the topics of the articles a template can be used in. Many cleanup templates can only be applied to articles from certain subject areas. An example with a particularly obvious restriction is the template *in-universe* (see Table 1), which should only be applied to articles about fiction. This topical restriction is neither explicitly defined nor automatically enforced, but it plays an important role in the quality flaw recognition task, as the remainder of this paper will show. While flaws merely concerning the structural or linguistic properties of an article are less restricted to individual topics, they are still affected by a certain degree of *topical preference*. Many subject areas in Wikipedia are organized in *WikiProjects*[8], which have their own ways of reviewing and ensuring quality within their topical scope. Depending on the quality assurance processes established in a WikiProject, different importance is given to individual types of flaws. Thus, the distribution of cleanup templates regarding structural or grammatical flaws is also biased towards certain topics.

[8]`http://en.wikipedia.org/wiki/WP:PROJ`

We will henceforth subsume the concept of topical preference under the term topical restriction.

Quality Flaw Recognition Based on the above definition of quality flaws, we define the quality flaw recognition task similar to Anderka et al. (2012) as follows: Given a sample of articles in which each article has been tagged with any cleanup template τ_i from a specific template cluster T_f thus marking all articles in the sample with a quality flaw f, it has to be decided whether or not an unseen article suffers from f.

4 Data Selection and Corpus Creation

For creating our corpora, we start with selecting all cleanup templates listed under the categories *neutrality* and *style* in the typology of cleanup templates provided by Anderka and Stein (2012a). Each of the selected templates serves as the nucleus of a template cluster that potentially represents a quality flaw. To each cluster, we add all templates that are synonymous to the nucleus. The synonyms are listed in the template description under *redirects* or *shortcuts*. Then we iteratively add all synonyms of the newly added template until no more redirects can be found. Furthermore, we manually inspect the lists of similar templates in the *see also* sections of the template descriptions and include all templates that refer to the same concept as the other templates in the cluster. As mentioned earlier, this is a subjective task and largely depends on the desired granularity of the flaw definitions. We finally merge semantically similar template clusters to avoid too fine grained flaw distinctions.

As a result, we obtain a total number of 94 template clusters representing 60 style flaws and 34 neutrality flaws. From each of these clusters, we remove templates with inline or section scope due to the reasons outlined in Section 3. We also remove all templates that are restricted to pages other than articles (e.g. discussion or user pages).

We use the Java Wikipedia Library (Zesch et al., 2008) to extract all articles marked with the selected templates. We only regard flaws with at least 500 affected articles in the snapshot of the English Wikipedia from January 4, 2012. Table 1 lists the final sets of flaws used in this work. For each flaw, the nucleus of the template cluster is provided along with a description, the number of affected articles, and the cluster size. We make the corpora freely available for download under `http://www.ukp.tu-darmstadt.de/data/wiki-flaws/`.

Flaw	κ	F_1
Advert	.60	.80
Confusing	.60	.80
Copy-edit	.00	.50
Essay-like	.60	.80
Globalize:	.60	.80
In-universe	.80	.90
Peacock	.70	.84
POV	.60	.80
Technical	.90	.95
Tone	.40	.70
Trivia	.20	.60
Weasel	.50	.74

Table 2: Agreement of human annotator with gold standard

Agreement with Human Rater

Quality flaw detection in Wikipedia is based on the assuption that cleanup templates are valid markers of quality flaws. In order to test the reliability of these user assigned templates as quality flaw markers, we carried out an annotation study in which a human annotator was asked to perform the binary flaw detection task manually. Even though the human performance does not necessarily provide an upper boundary for the automatic classification task, it gives insights into potentially problematic cases and ill-defined annotations. The annotator was provided with the template definitions from the respective template information page as instructions. For each of the 12 article scope flaws, we extracted the plain text of 10 random flawed articles and 10 random untagged articles. The annotator had to decide for each flaw individually whether a given text belonged to a flawed article or not. She was not informed about the ratio of flawed to untagged articles.

Table 2 lists the chance corrected agreement (Cohen's κ) along with the F_1 performance of the human annotations against the gold standard corpus. The templates *copy-edit* and *trivia* yielded the lowest performance in the study. Even though *copy-edit* templates are assigned to whole articles, they refer to grammatical and stylistic problems of relatively small portions of the text. This increases the risk of overlooking a problematic span of text, especially in longer articles. The *trivia* template, on the other hand, designates sections that contain miscellaneous information that are not well integrated in the article. Upon manual inspection, we found a wide range of possible manifestations of

this flaw ranging from an agglomeration of incoherent factoids to well-structured sections that did not exactly match the focus of the article, which is the main reason for the low agreement.

5 Selection of Reliable Training Instances

Independent from the classification approach used to identify flawed articles, reliable training data is the most important prerequisite for good predictions. On the one hand, we need a set of examples that reliably represent a particular flaw, while on the other hand, we need counterexamples which reliably represent articles that do not suffer from the same flaw. The latter aspect is most important for discriminative classification approaches, since they rely on negative instances for training the classifier. However, reliable negative instances are also important for one-class classification approaches, since it is only for the counterexamples (or outliers) that the performance of one-class classifiers can be sufficiently evaluated. It is furthermore important that the positive and the negative instances do not differ *systematically* in any respect other than the presence or absence of the respective flaws, since any systematic difference will bias the classifier. In this context, the topical restrictions of cleanup templates have to be taken into account. In the following, we describe our approach to extracting reliable training instances from the quality flaw corpora.

5.1 Reliable Positives

In previous work, the latest available versions of flawed articles have been used as positive training instances. However, we found upon manual inspection of the data that a substantial number of articles has been significantly edited between the time t_τ, at which the template was first assigned, and the time t_e, at which the articles have been extracted. Using the latest version at time t_e can thus include articles in which the respective flaw has already been fixed without removing the cleanup template. Therefore, we use the revision of the article at time t_τ to assure that the flaw is still present in the training instance.

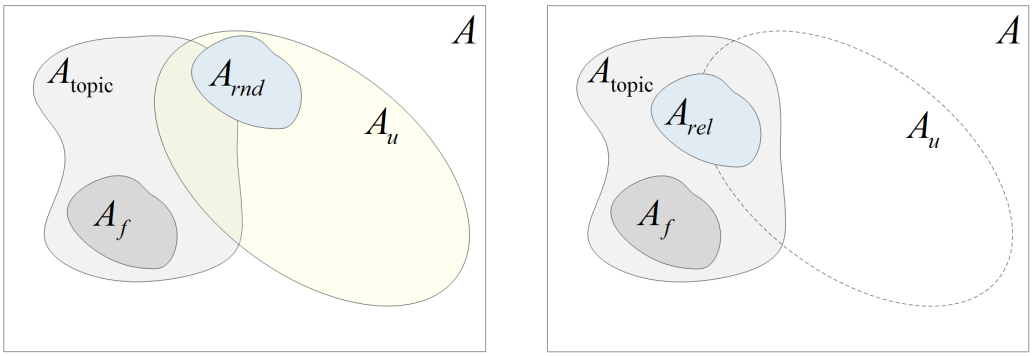
We use the Wikipedia Revision Toolkit (Ferschke et al., 2011), an enhancement of the Java Wikipedia Library, to gain access to the revision history of each article. For every article in the corpus of positive examples for flaw f that is marked with template $\tau \in T_f$, we backtrack the revision history chronologically, until we find the first revision $r_{t_\tau - 1}$ that is not tagged with τ . We then add the succeeding revision r_{t_τ} to the corpus of reliable positives for flaw f. In Section 6, we show that the classification performance improves for most flaws when using reliable positives instead of the latest available article versions.

5.2 Reliable Negatives and Topical Restriction

A central problem of the quality flaw recognition approach is the fact that there are no articles available that are tagged to not contain a particular quality problem. So far, two solutions to this issue have been proposed in related work. Anderka et al. (2012) tackle the problem with a one-class classifier that is trained on the positive instances alone thus eradicating the need for negative instances in the training phase. However, in order to evaluate the classifier, a set of outliers is needed. The authors circumvent this issue by evaluating their classifiers on a set of random untagged instances and a set of featured articles and argue that the actual performance of predicting the quality flaws lies between the two.

Ferretti et al. (2012) follow a two step classification approach (PU learning) that first uses a Naive Bayes classifier trained on positive instances and random untagged articles to pre-classify the data. In a second phase, they use the negatives identified by the Naive Bayes classifier to train a Support Vector Machine that produces the final predictions. Even though the Naive Bayes classifier was supposed to identify reliable negatives, the authors found no significant improvement over a random selection of negative instances, which effectively renders the PU learning approach redundant.

None of the above approaches consider the issue of topical restriction mentioned in Section 3, which introduces a systematic bias to the data. Both approaches sample random negative instances A_{rnd} for any given set of flawed articles A_f from a set of untagged articles A_u (see Fig. 1a). In order to factor out the article topics as a major characteristic for distinguishing flawed articles from the set of outliers, reliable negative instances A_{rel} have to be sampled from the restricted topic set A_{topic} that contains articles with a topic distribution similar to the flawed articles in A_f (see Fig. 1b). This will avoid the systematic bias and



(a) Random negatives

(b) Reliable negatives

Figure 1: Sampling of negative instances for a given set of flawed articles (A_f). Random negatives (A_{rnd}) are sampled from articles without any cleanup templates (A_u). Reliable negatives (A_{rel}) are sampled from the set of articles (A_{topic}) with the same topic distribution as A_f

result in a more realistic performance evaluation.

In the following, we present our approach to extracting reliable negative training instances that conform with the topical restrictions of the cleanup templates. Without loss of generality, we assume that an article, from which a cleanup template $\tau \in T_f$ is deleted at a point in time d_τ, the article no longer suffers from flaw f at that point in time. Thus, the revision r_{d_τ} is a reliable *negative instance* for the flaw f. Additionally, since the article was once tagged with $\tau \in T_f$, it belongs to the the same restricted topic set A_{topic} as the positive instances for flaw f.

We use the *Apache Hadoop*[9] framework and *WikiHadoop*[10], an input format for Wikipedia XML dumps, for crawling the whole revision history of the English Wikipedia on a compute cluster. WikiHadoop allows each Hadoop mapper to receive adjacent revision pairs, which makes it possible to compare the changes made from one revision to the next. For every template $\tau \in T_f$, we extract all adjacent revision pairs $(r_{d_\tau - 1}, r_{d_\tau})$, in which the first revision contains τ and the second does not contain τ. Since there are occasions in which a template is replaced by another template from the same cluster, we ensure that r_{d_τ} does also not contain any other template from cluster T_f before we finally add the revision to the set of reliable negatives for flaw f.

In the remainder of this section, we evaluate the topical similarity between the positive and the negative set of articles for each flaw using both our method and the original approach. In Wikipedia, the topic of an article is captured by the categories assigned to it. In order to compare two sets of articles with respect to their topical similarity, we represent each article set as a category frequency vector. Formally, we calculate for each set the vector $\vec{C} = (w_{c_1}, w_{c_2}, \ldots, w_{c_n})$ with w_{c_i} being the weight of category c_i, i.e. the number of times it occurs in the set, and n being the total number of categories in Wikipedia. We can then estimate the topical similarity of two article sets by calculating the cosine similarity of their category frequency vectors $\vec{C_1} := A$ and $\vec{C_2} := B$ as

$$sim(A, B) = \frac{A \cdot B}{||A||\ ||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

Table 3 gives an overview of the similarity scores between each positive training set and the corresponding reliable negative set as well as between each positive set and a random set of untagged articles. We can see that the topics of articles in the positive training sets are highly similar to the topics of the corresponding reliable negative articles while they show little similarity to the articles in the random set. This implies that the systematic bias introduced by the topical restriction has largely been eradicated by our approach.

Individual flaws have differently strong topical restrictions. The strength of this restriction depends on the size of A_{topic}. That is, a flaw such as *in-universe* is restricted to a very narrow selection of articles, while a flaw such as *copy edit* can be applied to most articles and rather shows a topical preference due to reasons outlined in Section 3. It

[9] http://hadoop.apache.org

[10] https://github.com/whym/wikihadoop

Flaw	Cosine Similarity (A_f, A_{rel})	(A_f, A_{rnd})
Advert	.996	.118
Confusing	.996	.084
Copy-edit	.993	.197
Essay-like	.996	.132
Globalize	.992	.023
In-universe	.996	.014
Peacock	.995	.310
POV	.994	.252
Technical	.995	.018
Tone	.996	.228
Trivia	.980	.184
Weasel	.976	.252

Table 3: Cosine similarity scores between the category frequency vectors of the flawed article sets and the respective random or reliable negatives

is therefore to be expected that that flaws with a small A_{topic} are more prone to the topic bias.

6 Experiments

In the following, we describe our system architecture and the setup of our experiments. Our system for quality flaw detection follows the approach by Ferschke et al. (2012b), since it has been particularly designed as a modular system based on the Unstructured Information Management Architecture[11], which makes it easy to extend. Instead of using *Mallet* (McCallum, 2002) as a machine learning toolkit, we employ the *Weka Data Mining Software* (Hall et al., 2009) for classification, since it offers a wider range of state-of-the-art machine learning algorithms. For each of the 12 quality flaws, we employ three different dataset configurations. The `BASE` configuration uses the newest version of each flawed article as positive instances and a random set of untagged articles as negative instances. The `RELP` configuration uses reliable positives, as described in Section 5.1, in combination with random outliers. Finally, the `RELALL` configuration employs reliable positives in combination with the respective reliable negatives as described in Section 5.2.

Features

An extensive survey of features for quality flaw recognition has been provided by Anderka et al. (2012). We selected a subset of these features for our experiments and grouped them into four feature sets in order to determine how well different combinations of features perform in the task.

[11] `http://uima.apache.org`

Category	Feature type	NONGRAM	NGRAM	NOWIKI	ALL
Lexical	Article ngrams		•	•	•
	Info to noise ratio	•		•	•
Network	# External links	•			•
	# Outlinks	•			•
	# Outlinks per sentence	•			•
	# Language links	•			•
References	Has reference list	•			•
	# References	•			•
	# References per sentence	•			•
Revision	# Revisions	•			•
	# Unique contributors	•			•
Structure	# Empty sections	•			•
	Mean section size	•			•
	# Sections	•			•
	# Lists	•			•
	Question rate	•		•	•
Readability	ARI	•		•	•
	Coleman-Liau	•		•	•
	Flesch	•		•	•
	Flesch-Kincaid	•		•	•
	Gunning Fog	•		•	•
	Lix	•		•	•
	SMOG-Grading	•		•	•
Named Entity	# Person entities*	•		•	•
	# Organization entities*	•		•	•
	# Location entities*	•		•	•
Misc	# Characters	•		•	•
	# Sentences	•		•	•
	# Tokens	•		•	•
	Average sentence length	•		•	•
	Article lead length	•			•
	Lead to article ratio	•			•
	# Discussions	•			•

* newly introduced feature
# number of instances

Table 4: Feature sets used in the experiments

Table 4 lists all feature types used in our experiments.

Since the feature space becomes large due to the ngram features, we prune it in two steps. First, we filter the ngrams according to their document frequency in the training corpus. We discard all ngrams that occur in less than x% and more than y% of all documents. Several values for x and y have been evaluated in parameter tuning experiments. The best results have been achieved with x=2 and y=90. In a second step, we apply the Information Gain feature selection approach (Mitchell, 1997) to the remaining set to determine the most useful features.

Learning Algorithms

We evaluated several learning algorithms from the Weka toolkit with respect to their performance on

Algorithm	Average F_1
SVM RBF Kernel	**0.82**
AdaBoost (decision stumps)	0.80
SVM Poly Kernel	0.79
RBF Network	0.78
SVM Linear Kernel	0.77
SVM PUK Kernel	0.76
J48	0.75
Naive Bayes	0.72
MultiBoostAB (decision stumps)	0.71
Logistic Regression	0.60
LibSVM One Class	0.67

Table 5: Average F_1-scores over all flaws on RELP using all features

the quality flaw recognition task. Table 5 shows the average F_1-score of each algorithm on the RELP dataset using all features. The performance has been evaluated with 10-fold cross validation on 2,000 documents split equally into positive and negative instances. One class classifiers are trained on the positive instances alone. We determined the best parameters for each algorithms in a parameter optimization run and list the results of the best configuration.

Overall, Support Vector Machines with RBF kernels yielded the best average results and outperformed the other algorithms on every flaw. We used a sequential minimal optimization (SMO) algorithm (Platt, 1998) to train the SVMs and used different γ-values for the RBF kernel function. In contrast to Ferretti et al. (2012), we did not see significant improvements when optimizing γ for each individual flaw, so we determined one best setting for each dataset. Since SVMs with RBF kernels are a special case of RBF networks that fit a single basis function to the data, we also used general RBF networks that can employ multiple basis functions, but we did not achieve better results with that approach.

One-class classification, as proposed by Anderka et al. (2012), did not perform well within our setup. Even though we used an out-of-the-box one class classifier, we achieve similar results as Anderka et al. in their pessimistic setting, which best resembles our configuration. However, the performance still lacks behind the other approaches in our experiments. The best performing algorithm reported by Ferschke et al. (2012b), AdaBoost with decision stumps as a weak learner, showed the second best results in our experiments.

7 Evaluation and Discussion

The SVMs achieve a similar cross-validated performance on all feature sets containing ngrams, showing only minor improvements for individual flaws when adding non-lexical features. This suggests that the classifiers largely depend on the ngrams and that other features do not contribute significantly to the classification performance. While structural quality flaws can be well captured by special purpose features or intensional modeling, as related work has shown, more subtle content flaws such as the neutrality and style flaws are mainly captured by the wording itself. Textual features beyond the ngram level, such as syntactic and semantic qualities of the text, could further improve the classification performance of these flaws and should be addressed in future work. Table 6 shows the performance of the SVMs with RBF kernel[12] on each dataset using the *NGRAM* feature set. The average performance based on *NOWIKI* is slightly lower while using *ALL* features results in slightly higher average F_1-scores. However, the differences are not statistically significant and thus omitted. Classifiers using the *NONGRAM* feature set achieved average F_1-scores below 0.50 on all datasets. The results have been obtained by 10-fold cross validation on 2,000 documents per flaw.

The classifiers trained on reliable positives and random untagged articles (RELP) outperform the respective classifiers based on the BASE dataset for most flaws. This confirms our original hypothesis that using the appropriate revision of each tagged article is superior to using the latest available version from the dump. The performance on the RELALL dataset, in which the topic bias has been factored out, yields lower F_1-scores than the two other approaches. Flaws that are restricted to a very narrow set of topics (i.e. A_{topic} in Fig. 1b is small), such as the *in-universe* flaw, show the biggest drop in performance. Since the topic bias plays a major role in the quality flaw detection task, as we have shown earlier, the topic-controlled classifier cannot take advantage of the topic information, while the classifiers trained on the other corpora can make use of these characteristic as the most discriminative features. In the RELALL setting, however, the differences between the positive and negative instances are largely determined by the flaws alone. Classifiers trained on

[12] γ=0.01 for BASE,RELP and γ=0.001 for RELALL

such a dataset therefore come closer to recognizing the actual quality flaws, which makes them more useful in a practical setting despite lower cross-validated scores.

In addition to cross-validation, we performed a cross-corpus evaluation of the classifiers for each flaw. Therefore, we evaluated the performance of the unbiased classifiers (trained on RELALL) on the biased data (RELP) and vice versa. Hereby, the positive training and test instances remain the same in both settings, while the unbiased data contains negative instances sampled from A_{rel} and the unbiased data from A_{rnd} (see Figure 1). With the *NGRAM* feature set, the reliable classifiers outperformed the unreliable classifiers on all flaws that can be well identified with lexical cues, such as *Advert* or *Technical*. In the biased case, we found both topic related and flaw specific ngrams among the most highly ranked ngram features. In the unbiased case, most of the informative ngrams were flaw specific expressions. Consequently, biased classifiers fail on the unbiased dataset in which the positive and negative class are sampled from the same topics, which renders the highly ranked topic ngrams unusable. Flaws that do not largely rely on lexical cues, however, cannot be predicted more reliably with the unbiased classifier. This means that additional features are needed to describe these flaw. We tested this hypothesis by using the full feature set *ALL* and saw a substantial improvement on the side of the unbiased classifier, while the performance of the biased classifier remained unchanged.

A direct comparison of our results to related work is difficult, since neutrality and style flaws have not been targeted before in a similar manner. However, the *Advert* flaw was also part of the ten flaw types in the PAN Quality Flaw Recognition Task (Anderka and Stein, 2012b). The best system achieved an F_1 score of 0.839, which is just below the results of our system on the BASE dataset, which is similar to the PAN setup.

Flaw	BASE	RELP	RELALL
Advert	.86	.88	.75
Confusing	.76	.80	.70
Copy edit	.81	.73	.72
Essay-like	.79	.83	.64
Globalize	.85	.87	.69
In-universe	.96	.96	.69
Peacock	.77	.82	.69
POV	.75	.80	.71
Technical	.87	.88	.67
Tone	.70	.79	.69
Trivia	.72	.77	.70
Weasel	.69	.77	.72
⌀	.79	.83	.70

Table 6: F_1 scores for the 10-fold cross validation of the SVMs with RBF kernel on all datasets using *NGRAM* features

8 Conclusions

We showed that text classification based on Wikipedia cleanup templates is prone to a topic bias which causes skewed classifiers and overly optimistic cross-validated evaluation results. This bias is known from other text classification applications, such as authorship attribution, genre detection and native language detection. We demonstrated how to avoid the topic bias when creating quality flaw corpora. Unbiased corpora are not only necessary for training unbiased classifiers, they are also invaluable resources for gaining a deeper understanding of the linguistic properties of the flaws. Unbiased classifiers reflect much better the performance of quality flaw recognition "in the wild", because they detect actual flawed articles rather than identifying the articles that are prone to certain quality due to their topic or subject matter. In our experiments, we presented a system for identifying Wikipedia articles with style and neutrality flaws, a novel category of quality problems that is of particular importance within and outside of Wikipedia. We showed that selecting a reliable set of positive training instances mined from the revision history improves the classification performance. In future work, we aim to extend our quality flaw detection system to not only find articles that contain a particular flaw, but also to identify the flaws within the articles, which can be achieved by leveraging the positional information of in-line cleanup templates.

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806, and by the Hessian research excellence program "Landes-Offensive zur Entwicklung Wissenschaftlich-Ökonomischer Exzellenz" (*LOEWE*) as part of the research center "Digital Humanities".

References

Maik Anderka and Benno Stein. 2012a. A Breakdown of Quality Flaws in Wikipedia. In *2nd Joint WICOW/AIRWeb Workshop on Web Quality*, pages 11–18, Lyon, France.

Maik Anderka and Benno Stein. 2012b. Overview of the 1st International Competition on Quality Flaw Prediction in Wikipedia. In *CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers*.

Maik Anderka, Benno Stein, and Nedim Lipka. 2012. Predicting Quality Flaws in User-generated Content: The Case of Wikipedia. In *35th International ACM Conference on Research and Development in Information Retrieval*, Portland, OR, USA.

Julian Brooke and Graeme Hirst. 2011. Native language detection with 'cheap' learner corpora. In *Learner Corpus Research 2011 (LCR 2011)*.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10: Shared Task, pages 1–12, Stroudsburg, PA, USA. Association for Computational Linguistics.

Edgardo Ferretti, Donato Hernández Fusilier, Rafael Guzmán-Cabrera, Manuel Montes y Gómez, Marcelo Errecalde, and Paolo Rosso. 2012. On the Use of PU Learning for Quality Flaw Prediction in Wikipedia. In *CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers*.

Oliver Ferschke, Torsten Zesch, and Iryna Gurevych. 2011. Wikipedia Revision Toolkit: Efficiently Accessing Wikipedia's Edit History. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. System Demonstrations*, pages 97–102, Portland, OR, USA.

Oliver Ferschke, Iryna Gurevych, and Yevgen Chebotar. 2012a. Behind the Article: Recognizing Dialog Acts in Wikipedia Talk Pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 777–786, Avignon, France.

Oliver Ferschke, Iryna Gurevych, and Marc Rittberger. 2012b. FlawFinder: A Modular System for Predicting Quality Flaws in Wikipedia. In *CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers*, Rome, Italy.

Aidan Finn and Nicholas Kushmerick. 2006. Learning to classify documents according to genre. *Journal of the American Society for Information Science and Technology*, 57(11):1506–1518.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1):10–18.

Moshe Koppel and Jonathan Schler. 2003. Exploiting stylistic idiosyncrasies for authorship attribution. In *Workshop on Computational Approaches to Style Analysis and Synthesis*, pages 69–72.

K. Luyckx and W. Daelemans. 2005. Shallow text analysis and machine learning for authorship attribution. In *Proceedings of the Fifteenth Meeting of Computational Linguistics in the Netherlands (CLIN 2004)*, pages 149–160.

Andrew Kachites McCallum. 2002. MALLET: A Machine Learning for Language Toolkit.

George K. Mikros and Eleni K. Argiri. 2007. Investigating topic influence in authorship attribution. In *Proceedings of the SIGIR 2007 International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection, PAN 2007*, Amsterdam, Netherlands.

Thomas Mitchell. 1997. *Machine Learning*. McGraw-Hill Education, New York, NY, USA, 1st edition.

John C Platt. 1998. Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods: Support Vector Learning*, pages 185–208, Cambridge, MA, USA.

Besiki Stvilia, Michael B. Twidale, Linda C. Smith, and Les Gasser. 2008. Information Quality Work Organization in Wikipedia. *Journal of the American Society for Information Science and Technology*, 59(6):983–1001.

György Szarvas, Veronika Vincze, Richárd Farkas, György Móra, and Iryna Gurevych. 2012. Cross-genre and cross-domain detection of semantic uncertainty. *Comput. Linguist.*, 38(2):335–367.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco.