Automated Decision Support for Human Tasks in a Collaborative System: The Case of Deletion in Wikipedia

Bluma S. Gelley Torsten Suel

Polytechnic Institute of NYU
2 Metrotech Center
Brooklyn, NY 11201
bgelley@nyu.edu, suel@poly.edu

ABSTRACT

Wikipedia's low barriers to participation have the unintended effect of attracting a large number of articles whose topics do not meet Wikipedia's inclusion standards. Many are quickly deleted, often causing their creators to stop contributing to the site. We collect and make available several datasets of deleted articles, heretofore inaccessible, and use them to create a model that can predict with high precision whether or not an article will be deleted. We report precision of 98.6% and recall of 97.5% in the best case and high precision with lower, but still useful, recall, in the most difficult case. We propose to deploy a system utilizing this model on Wikipedia as a set of decision-support tools to help article creators evaluate and improve their articles before posting, and new article patrollers make more informed decisions about which articles to delete and which to improve.

Categories and Subject Descriptors

H.5.3. Collaborative Computing; Computer Supported Collaborative Work

General Terms

Measurement, Performance, Human Factors,

Keywords

Wikipedia; Collaborative Systems; Deletion; Automating Human Tasks; Decision Support; Classification;

1. INTRODUCTION

Wikipedia, the largest encyclopedia ever and the sixth most visited website in the world [3], has not a single writer on staff. All the work of creating and improving the encyclopedia is performed by millions of volunteers from around the world. The wiki software the site is built on allows anyone to add to Wikipedia by just clicking on an edit button – no account required. Many of these volunteers contribute just one edit, while others make tens of thousands. By minimizing barriers to participation, Wikipedia has successfully encouraged many people with valuable knowledge to share it with the world, and has become an important information resource.

In this paper, we focus on the English Wikipedia, since it is by far the largest; it currently has more than 4.2 million articles, over 30 times more than the online Encyclopedia Britannica [40][1]. The ease with which users can create new articles is one major reason for Wikipedia's impressive subject coverage.

It also means, however, that it is very easy to create articles that are not worthy of inclusion in Wikipedia, such as those consisting solely of personal attacks or advertising, or articles about people or other entities that are not noteworthy enough to warrant an encyclopedia article. Though the question of what, exactly, is worthy of inclusion in Wikipedia is a controversial issue in the Wikipedia community, a general consensus has been arrived at as to what makes a topic encyclopedic, or **notable***,* enough to be included in the encyclopedia. Generally, this means "extensive" coverage of the topic in secondary sources, such as books or the media [37]; specific guidelines cover the notability of people, books, events, and many other categories.

Despite these rules, and the additional safeguard that only registered users can create new pages, hundreds of articles are created on Wikipedia each day that are not encyclopedic. These range from the obviously promotional to the absurd. A sampling of articles we collected, all created on a single day, includes pages about "a [sic] Iowa business leader and a political activist" whose major accomplishment seems to have been owning a pork farm, a car-washing company created by a teenager in "Cedar Way" (no other location given) over his summer vacation (now defunct), and a description of 'a sort of mythological race' composed of trombones whose goal is to 'survive and evolve into a superior race.' Several hundred articles like this are created every day. Articles like these, and any others that do not meet the notability standards, are eligible for deletion, or removal from the site, using the process described in Section 2.

In recent years, some have noticed a troubling trend: Wikipedia is finding it more difficult to retain new contributors. While there is a steady influx of new registrations, very few go on to become active editors, and the active editor cohort size is actually decreasing [10, 25]. This is a serious problem – Wikipedia cannot exist without a critical mass of active contributors – and has been extensively studied [10, 13, 15, 16, 27]. One possible cause for this attrition is the aggressiveness with which newly created pages are treated. New editors who leave after a short period of time often cite the speed and lack of explanation with which their

WikiSym'13, Aug 05-07 2013, Hong Kong, China.
Copyright 2013 ACM 978-1-4503-1852-5/13/08…$15.00.

[1] In this paper, we use footnotes to refer readers to resources that we have used and references to cite sources for assertions made; both can point to Wikipedia pages.

articles were deleted as a cause of their defection [11]. Indeed, new editors whose edits were reverted by more experienced ones are less likely to continue contributing [15]; among users whose first edit created an article, those whose articles were deleted were six times more likely to abandon Wikipedia immediately than those whose articles were not [27].

The Wikipedians who vet new articles, while acknowledging that over-zealous article deletion is a problem, argue that a large percentage of new articles are so clearly inappropriate that quick deletions are not only warranted in many cases, but are necessary to ensure that the site is not overrun with promotional and other unencyclopedic pages [38].

1.1 Previous Work

To shed some light on this phenomenon, we previously [14] conducted an in-depth investigation of Speedy Deletion, the most common form of deletion on Wikipedia. We collected a dataset of deleted articles and compared several dozen features of each page to those of Wikipedia articles that had not been deleted. We also recorded the length of time between the creation of the article and its nomination for deletion. We found that rapid deletions were indeed widespread: 47% of Speedy Deleted articles were tagged for deletion within 10 minutes of their creation, often before more than a few lines of text were added. We also found, however, that there were clear differences between kept and Speedy Deleted articles, over metrics as varied as the number of references and the number of nouns in the article. These differences lessened, but remained significant, even when we accounted for the effect of the average difference in age between kept and deleted articles [14].

1.2 Overview and Contributions

In this paper, we attempt to model the differences (if any) between articles that fit the generally-accepted deletion criteria and those that do not, and to build a classifier that can predict which articles are likely to be deleted under current conditions. We use pages that were deleted and those that were kept on the site as approximations for these two classes. We extend our work beyond the single deletion method (Speedy Deletion) covered in [14] to include the two other major deletion forms, Proposed Deletion (PROD) and Deletion Discussion (AfD). Our focus is on articles that are likely to have been created in good faith but do not meet Wikipedia's standards for inclusion. We anticipate our model being used to find articles that meet the deletion criteria so they can potentially be improved, as a self-evaluation tool to help new article creators bring their articles up to basic Wikipedia standards before creation, or (with some caveats, see Section 8) as a decision support tool to help the Wikipedians who vet new articles make better decisions about which to delete.

The contributions of this work are as follows:

- An intelligent, high-precision system for predicting which articles will be deleted from Wikipedia. We experiment with the three primary forms of deletion and build individual classifiers for each one. These classifiers can be used for any or all of the possible uses mentioned above and detailed in Section 4 below.
- Several datasets totaling several thousand articles deleted from Wikipedia, along with over 15,000 which were kept. This data is available for download[2].

The remainder of this paper is organized as follows. Section 2 contains some background and statistics on the various forms of deletion used in Wikipedia, and Section 3 explores related work in the area. Section 4 presents some anticipated uses for our model. Our data, including collection and processing methods, is described in Section 5. Features are discussed in Section 6, and the actual evaluation in Section 7. Section 8 discusses the limitations of this work and possible implementation concerns, and Section 9 concludes and sets out some potential future work.

2. BACKGROUND AND STATISTICS

There are several modes of article deletion on Wikipedia. Here, we give a brief description of each and some statistics about their frequency. (For a flowchart detailing the deletion process, see [24].) To prevent abuse, only Wikipedia superusers, known as administrators, can delete pages, but any user can nominate an article for deletion by adding a simple template to the page (see Figure 1). Deletion is supposed to be a collective decision, but a large number of articles are so clearly unencyclopedic that requiring a long discussion before deleting each one would place an undue burden on the community. The Criteria for Speedy Deletion [36] outline several dozen reasons an article can be deleted without discussion. Among these are copyright violation, unambiguous advertising, and personal attacks. Once nominated, the articles are listed on a central page and can be unilaterally deleted by any administrator.

Articles which are uncontroversially out of place on Wikipedia, but do not meet any of the Speedy Deletion criteria, can be Proposed for Deletion, or PRODed. PRODs are given a waiting period of a week; if anyone, even the article's creator, contests the nomination, the article is de-PRODed; otherwise, it is deleted. There is also a separate, 10-day PROD process for unattributed biographies of living people. During the time of our data collection (Oct. – Dec. 2011), we found that between 30 and 70 articles were PRODed every day, with a mean of 45 and median of 44.

Deletion Discussion is the most intensive method of deletion. Any article whose deletion might be at all controversial must be brought to the Articles for Deletion (AfD) page for community discussion. Anyone can express an opinion as to whether it should be kept or deleted, along with an explanation of their position. After at least a week, an administrator who was not involved in the discussion determines what the general consensus (not based on majority voting) was and carries out the appropriate action. We found that the number of new deletion discussions opened each day from January 1 – March 31, 2013 ranged from ~30 to over 100, with a mean of 60 and median of 59.

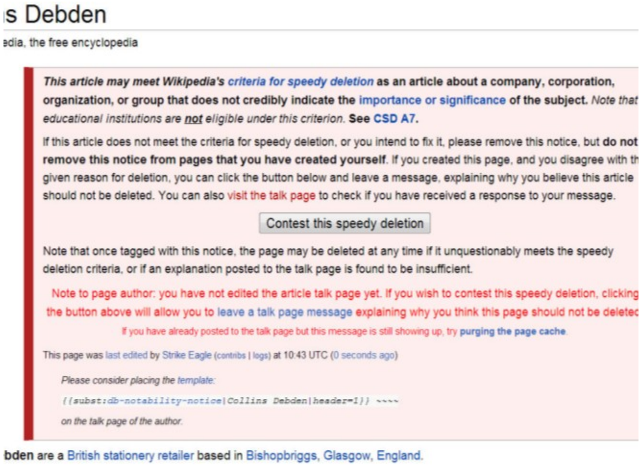
While Deletion Discussion is officially the default deletion method, there are many times more Speedy Deletions than AfD's: several hundred (measured throughout February 2013), as compared to ~60 (see above), daily deletions. Articles nominated for AfD have about a 50% deletion rate, while more than 70% of Speedy nominees are deleted. PRODs have an even higher deletion rate, averaging 86%. (We calculated AfD deletion rates for all deletion discussions opened between January 1 and March 31, 2013. The Speedy and PROD deletion rates were calculated for our Speedy Deletion and PROD datasets, collected from Oct.-Dec. 2011.)

There are 22 Speedy Deletion criteria that can be applied to articles. Of these, several can be classified as technical reasons, such as pages created in error, pages that must be deleted to make

[2] https://github.com/bsgelley/Wikipedia-deletion-data.git.

way for a move, or pages dependent on a non-existent page. These comprise about 12% [12] of all Speedy Deletion nominations. The rest of the criteria include pages containing only vandalism, attack pages, blatant hoaxes, and articles whose topics are clearly too insignificant to be included in Wikipedia.

Articles that are deleted are removed from the site and are inaccessible to the public. The only record is maintained in the deletion log, which contains information about the deletion, but not the article itself. This fact has impacted all attempts to study deletion until now, since the content of deleted articles was not available to researchers.



s Debden

edia, the free encyclopedia

This article may meet Wikipedia's criteria for speedy deletion **as an article about a company, corporation, organization, or group that does not credibly indicate the importance or significance of the subject.** *Note that educational institutions are* ***not*** *eligible under this criterion.* **See CSD A7.**

If this article does not meet the criteria for speedy deletion, or you intend to fix it, please remove this notice, but **do not remove this notice from pages that you have created yourself.** If you created this page, and you disagree with th given reason for deletion, you can click the button below and leave a message, explaining why you believe this article should not be deleted. You can also visit the talk page to check if you have received a response to your message.

Contest this speedy deletion

Note that once tagged with this notice, the page may be deleted at any time if it unquestionably meets the speedy deletion criteria, or if an explanation posted to the talk page is found to be insufficient.

Note to page author: you have not edited the article talk page yet. If you wish to contest this speedy deletion, clicking the button above will allow you to leave a talk page message explaining why you think this page should not be deletec

If you have already posted to the talk page but this message is still showing up, try **purging the page cache.**

This page was last edited by Strike Eagle (contribs | logs) at 10:43 UTC (0 seconds ago)

Please consider placing the template:

```
{{subst:db-notability-notice|Collins Debden|header=1}} ~~~~
```

on the talk page of the author.

bden are a British stationery retailer based in Bishopbriggs, Glasgow, England.

Figure 1. An article nominated for Speedy Deletion. Note the deletion template at the top of the page.

3. RELATED WORK

There has been extensive work on the subject of detecting and removing vandalism in Wikipedia, mostly using machine learning; for example, [22, 21, 1, 29, 8, 23, 28], and [2]. At least two of these ([8] and [29]) have been deployed autonomously or semi-autonomously on Wikipedia to aid humans in fighting vandalism. None of these solutions address the issue of unencyclopedic pages as a problem separate from vandalism.

Not much work, conversely, has been done on deletion in Wikipedia, particularly in the area of detecting pages to be deleted. What little research has been done has focused primarily on Deletion Discussion, just one of four deletion methods and one that comprises less than 25% of all deletions (see Section 2). In [26], the authors analyze AfD discussions for signs of external influence in voting; in [20] the authors explore the effects that various characteristics of the groups participating in deletion discussions have on the decisions made. Schneider et al. [24] identify the most common decision factors used in deletion arguments, while [41] automatically summarizes deletion discussions to help administrators make decisions. In [12], the authors discuss the average number of participants in a deletion discussion and some characteristics of the users who tend to participate. They also briefly discuss Speedy Deletion, tallying the percentages of all Speedy Deletions carried out for each of the Speedy criteria. Lam and Reidl [19] also study Speedy Deletions, using a rough proxy for notability to attempt to determine whether articles deleted for lack of notability were actually non-notable. Revision deletion, a topic only somewhat related to article deletion, is discussed in [30]. No one has studied the content of the deleted pages themselves, however, since that was inaccessible. In our previous work [14], we undertake an intensive examination of deleted pages and their characteristics. Using the dataset and features described below, we compare the characteristics of Speedy Deleted pages to a set of pages that were not deleted.

In [4], the authors address a related, but different problem: predicting flaws in Wikipedia articles. One of the flaws they address is lack of notability, which is one of the most common reasons for deletion, and one which we focus on in this work. Due, however, to the nature of the difference between our problem and theirs, they are only able to compare to an extremely optimistic dataset with little relation to the real-world problem they face. When comparing to a more realistic dataset similar to our own, they achieve significantly worse results than we do. In practice, the problem of identifying articles that may be tagged for correction because of possible lack of notability is very different from our problem of identifying pages that should be deleted, and direct comparison of their work with ours is therefore impossible.

Finally, [7] attempts to model another Wikipedia process, that of choosing administrators. The issues that the authors have with modeling a complex, human-driven collaborative process and the limitations they acknowledge are very relevant to our problem.

4. ANTICIPATED USES

We envision three potential uses for our model.

1) Our model can be used in a tool to help experienced editors find pages that need improvement. Currently, tagging of pages that need work to be brought up to Wikipedia's standards is done manually. This is time-consuming and requires human editors to check every article, often multiple times over its life cycle. Our model can be used to predict which articles are likely to be deleted and are therefore probably of low quality, allowing editors to more efficiently find articles to work on. This can be deployed for two separate types of articles, newly created articles and older articles that may be of borderline encyclopedic quality. We therefore test our model on datasets that simulate both of these article types to ascertain whether it can effectively predict deletion in both of those cases.

2) Sometimes, no amount of improvement would make an article encyclopedic – its topic is simply not notable. As per Wikipedia's policy, these articles should be deleted. New Page Patrollers (NPPs), the Wikipedians who do the bulk of the work of finding articles for deletion and removing them, complain about the heavy burden this work places on them and the amount of time it takes [32]. Since the majority of NPPs also contribute significantly to the rest of the project [32], time spent patrolling new articles may be better spent working on improving the quality of the rest of Wikipedia. Our model can potentially be used as a decision-support tool to help NPPs quickly find articles that meet the standard deletion qualifications, reducing the amount of time and effort needed to review each page, and allowing them to focus on those pages that are of borderline encyclopedic quality and may be able to be improved. Some have claimed [32] that the huge volume of pages that must be reviewed makes them more likely to make too-hasty decisions to delete rather than giving each page a fair review; a tool that can make part of the process easier and faster may potentially reduce the number of articles deleted too hastily and make Wikipedia more welcoming to new article creators. As mentioned above, the model can also be used to find older articles that fit the deletion criteria; in this case, for deletion, if they are intrinsically unencyclopedic. Automated tools that flag

suspicious content and present it to human editors for review are heavily used on Wikipedia for vandalism detection; two examples are Huggle[3] and STiki [29]. This use of our model requires careful implementation, as some research has shown that automated tools actually increase the incidence of over-zealous deletions [16]; we discuss this and related concerns at the end of the paper.

3) Our model can also be used as a tool to help article creators evaluate their articles before creation. The information about what topics are suitable for Wikipedia that is offered to new editors when they attempt to create an article is well-written and clear, but is extremely long and detailed. Proper understanding of the inclusion criteria requires the reading of multiple other pages full of dense text and, in some cases, Wikipedia jargon. The Article Wizard[4], which tries to step users through the process of determining the suitability of their topics, is also heavily text-based and full of links to long, complicated policy pages. Even well-meaning newbies may be daunted by the denseness of the material, particularly those whose first language is not English [31]. It is easy to skip both using the Article Wizard and reading any guidelines and just create an article; judging by the number of articles that are deleted because they don't meet the guidelines, many new users do just that. Our model can be turned into a tool that will tell editors if their articles are likely to be deleted once created. This can help them decide whether to abandon the article, or perhaps improve it so that it meets the standards. We would like this tool to also include guidance for how to correctly write and source a Wikipedia article, as well as simple ways to let NPPs know that the article is being worked on and improved.

5. DATA

Unlike previous work, most of which focus on one form of deletion, we study all three of the main deletion methods; namely, Speedy Deletion (Speedy), Proposed Deletion (PROD), and Deletion Discussion (or Articles for Deletion – AfD). Each of these deletion methods is aimed at a different class of articles. We therefore build three separate classifiers, each of which detects articles that match the deletion criteria for each method.

5.1 Article Selection Criteria

In our examination of deleted articles, we found that many of them do not seem to have been created maliciously. The articles seem to be good-faith attempts by creators who are simply unaware of, or at most choose to ignore, Wikipedia's guidelines on what topics are appropriate for inclusion. In this sense, the deletion problem differs from vandalism, which by definition has malicious intent. We choose to focus on these good-faith articles because, unlike articles that are pure vandalism, these pages can seem quite similar to legitimate Wikipedia articles. It stands to reason, then, that automatically detecting them is a more difficult task than finding standard vandalism. These articles also require careful handling and must be treated differently than vandalism to avoid scaring away good-faith contributors. In detecting Speedy Deletion candidates, then, we avoid articles that can be found using current vandalism detection techniques or other heuristics; we limit ourselves to those Speedy Deletion criteria that are usually applied to articles created in good faith, but are not encyclopedic enough for Wikipedia. These criteria are "no indication of [the] importance" of their subjects, and "unambiguous advertising or promotion" [36]. These are the most common types of Speedy nominees, together comprising about 45% of all Speedy Deletions [12].

Since all articles that fall under the heading of vandalism are subject to Speedy Deletion only, the vast majority of PRODs and AfDs are good-faith creations. We therefore use all PRODs and AfDs without attempting to determine which were good-faith contributions.

As mentioned in Section 2, research on deletion has been hampered by the fact that deleted pages are no longer publicly available. Researchers, therefore, have been forced to rely on the limited data preserved in the deletion log, which concerns only the deletion itself and has no information about the page content. We therefore set out to collect a large number of pages that were likely to be deleted in the future before they could be deleted and lost. Since the procedure is different for each deletion method we study, we vary our data collection methods somewhat as well. The various datasets collected are summarized in Table 1. Next, we describe how these datasets were collected.

Table 1: Dataset Composition. Note that Original, Old, and New are all composed of Speedy Deleted articles.

Name	Kept Articles	Deleted Articles	Total
AfD	798	270	1068
PRODs	2036	991	3027
Original	1381	2444	3825
Old	580	191	771
New	2198	723	2921

5.2 Deletion Discussion (AfDs)

Articles nominated for Deletion Discussion (AfDs) are supposed to remain on Wikipedia for a least a week while the discussion is conducted, before being deleted. Each day's candidates are logged on a separate page. Once a week, we visited all Deletion Discussion log pages for that week and downloaded all listed pages. More than a week later (since some deletion discussions last longer than the stated week), we checked each page to determine whether it had been deleted. Since our aim is to detect a small number of pages for deletion among many other pages, we also downloaded an additional large number of articles that had been submitted for Deletion Discussion but not deleted, so the Kept class was significantly larger than the Deleted one. In this way, we approximated the class skew our classifier would encounter in the real world. This left us with a set of 270 articles that had been deleted, and 798 comparable pages that had been nominated for deletion, but kept, in February and March 2013.

5.3 Proposed Deletion (PROD)

We followed a similar procedure for PRODs, downloading a week's worth at once, then checking for deletion after each one's week-long waiting time (see Section 2) was up. Our final dataset is composed of 847 articles PRODed and deleted between October and December 2011, and an additional 141 from March 2013; these were added to ensure that any results obtained are not specific to the first short time period and will generalize to future

[3] http://en.wikipedia.org/wiki/Wikipedia:Huggle

[4] https://en.wikipedia.org/wiki/Wikipedia:Article_wizard

data as well. Articles that were nominated and not deleted were not included, for reasons explained in Section 5.4.5.

5.4 Speedy Deletion

5.4.1 Original Dataset

To obtain the deleted Speedies in our Original set, which are the most numerous and can be deleted at any time once nominated, we visited the Candidates for Speedy Deletion list page[5] once every 12 minutes from October to December 2011 and downloaded any new listings. We found this time span to be the optimal one for catching most of the listed pages before they were deleted; more frequent polling resulted in too few new pages at each iteration and caused an unnecessary load on Wikipedia's servers. We waited approximately a week (the vast majority of Speedies are deleted within that period [14]) and then checked which of the nominated pages had been deleted. As explained above (Section 5.1), we used only articles deleted for lack of significance, or for being advertising or promotion. We used the deletion nomination template placed on each page, which includes the reason for the nomination, to select pages deleted using these criteria. Articles with multiple deletion reasons were included if at least one was one of our selected criteria. We also filtered out non-article pages, such as files and Wikipedia policy pages. We were left with a set of 2444 articles that were Speedy- Deleted for the above-listed reasons. The comparison set of 1381 not-deleted (kept) articles included in the Speedy datasets were collected from Wikipedia using the process described in Section 5.4.4.

5.4.2 Old Dataset

One of our proposed use cases for the model includes detecting older articles that meet the deletion criteria but have managed to avoid deletion. To determine how effectively we can detect such articles, we create a set of all deleted articles in the Original (Speedy) dataset older than one week when nominated for deletion, as well as a random sample of 45% of all kept articles (to maintain a somewhat reasonable class skew) older than one week when downloaded. We refer to this dataset as "Old." This set is quite small, since the vast majority of Speedy Deletion nominees are nominated soon after they are created [14]. Still, good results on this dataset would be an encouraging sign that our model is effective for older articles as well as newer ones.

5.4.3 New Dataset

This set, which we refer to as New, contains approximately 15,000 articles downloaded from Wikipedia shortly after their creation in November and December 2012 (by downloading any new listings on the Special:NewPages page every 5 minutes); some of them were later deleted. We use a subset of these articles, those which were Speedy Deleted using our preferred criteria (See Section 5.1), in our experiments. This dataset closely simulates one of our goals: to monitor the stream of new articles added to Wikipedia and predict which ones will be deleted.

5.4.4 PROD and Speedy Deletion Comparison Sets

In this section, we describe the comparison sets of kept articles we use to complement the deleted articles in our PROD and Speedy (Original, Old, and New) sets. Unlike the AfDs, which naturally fall into two complementary classes, Kept and Deleted, our PROD and Speedy sets are composed only of deleted articles. This is because very few articles nominated for Speedy Deletion or PRODed end up being kept on Wikipedia; we have found that deletion rates for both methods are greater than 70% (see Section 2). Deletion rates for articles nominated using our chosen criteria are even higher, in the range of 80-90%. We are therefore unable to use articles nominated for deletion but not deleted as a comparison, as we do for AfDs. Instead, we choose a selection of articles from all over Wikipedia that were not deleted, and therefore, presumably, are sufficiently encyclopedic. These articles are included in the PROD and Speedy (Original, Old, and New) datasets as the Kept class. In creating the comparison sets, we attempt to minimize the effect of as many potential confounding factors as possible. One such factor is content. If the deleted articles were topically very different than the kept ones, the difference in the words used, or other semantic factors, might lead the classifier to incorrect conclusions. To find kept articles that matched the deleted ones topically, we utilize Wikipedia's category hierarchy. Wikipedia contains an exhaustive taxonomy for categorizing articles by topic[6], and most legitimate Wikipedia articles belong to at least one category. Fully 75% of the deleted articles in our dataset, on the other hand, did not belong to any. We therefore manually analyze a sampling of deleted articles to determine the most common topics among the articles in the deleted set. We then select the Wikipedia categories most similar to the chosen topics and use articles from those categories for the comparison set. In some cases, there is no Wikipedia category matching a common topic among the deleted pages. In those cases, we choose the Wikipedia category most similar to the given topic. We then take a random sampling of the articles in each chosen category. Since most of the deleted articles are fairly short (often due to having just been created when flagged for deletion), we include a large number of stubs (short articles; labeled as "stub" and maintained in separate categories) in the chosen topic areas, so article length should not be a deciding factor. We also remove any very long articles from the comparison set and leave a mixture of articles of different lengths. In all, the comparison set used for the Speedy experiments contains 1381 articles from 21 categories. For the PROD experiments, we use a superset of the Speedy comparison set, containing an additional 655 articles from 13 more categories, for a total of 2036 articles.

5.4.5 Comparison Sets - Possible Concerns

One issue with this approach is that there is no guarantee that the articles in our 'control' (Kept) sets are themselves worthy of inclusion in Wikipedia. While this concern may be true of some number of articles in our dataset, its large size and the random sampling methods employed to create it mean that the number of bad pages is unlikely to be statistically significant. In addition, Wikipedians heavily monitor new pages; the majority of deleted pages are flagged for deletion within hours, if not minutes, of their creation [33]. Few unencyclopedic pages are likely to have slipped through and remained on Wikipedia. Furthermore, we purposely chose articles for the comparison sets from the Wikipedia category hierarchy to reduce the likelihood of the articles being unencyclopedic. Since most articles are added to categories manually (or, if added by a bot, are usually part of an existing list of topics that are considered notable), those belonging to a category have generally had somewhat more oversight and are more likely to be encyclopedic than those that do not. Given these factors, we are confident that enough of our comparison set is

[5] http://en.wikipedia.org/wiki/Category:Candidates_for_speedy_d e letion

[6] http://en.wikipedia.org/wiki/Portal:Contents/Categories

encyclopedic enough to serve as a ground truth for classification. (This, of course, assumes that the deletion decisions made by Wikipedians are correct; see Section 8 for a discussion of this assumption and its possible pitfalls.)

A concern still remains about possible biases in the comparison set collection. We attempted to use articles as similar as possible to the deleted ones, differing only in that they were not deleted, and are therefore encyclopedic. Unfortunately, it is still possible that there are other, unknown, confounding factors that could bias our results.

A final concern is that many of the features we use, such as the number of revisions made to the article, are influenced by the age of the article. Since most deletion nominations are made early in the lifecycle of a page ([19]), most of the pages in the deleted class are much younger than those in the Kept class, and it is possible that many of the differences between the two classes can be explained as simply functions of the age of the page.

To address these last two issues, we experiment on the New dataset. It is the least biased of all the Speedy sets, since the comparison (Kept) articles have a natural relationship with the deleted ones and were not artificially selected (as the Original/Old comparison set was). Results on this dataset, then, are the most important to examine. If they are similar to the results achieved on the other Speedy sets (in particular, Original), it would provide strong corroboration that the Original comparison set is an appropriate one and is not too much affected by bias to be useful.

The New dataset also addresses the issue of possible bias caused by the difference in average article age between the Kept and Deleted classes. Since all the articles in the New set are of similar (very young) age, by experimenting on the New set, we effectively normalize by the age of the page and remove any confounding effect it might have. Given these benefits, it may seem reasonable to use only the New dataset. However, it does have some biases of its own. Since the articles are all very new, a classifier built using only this data might not generalize very well to older articles. We therefore experiment on both the New and Original datasets in the hope of obtaining useful information from both.

Finally, we also experiment on the Old dataset. Since these articles are all older, there is little difference in article age between the Kept and Deleted classes. Good results on this set, then, would further show that the age of the articles does not significantly affect the effectiveness of our model.

5.5 Preprocessing

We use Wikipedia's article export function[7] to download the articles as Wiki markup, wrapped in XML for easy parsing. We remove articles with no content (determined by file size). Of the feature classes listed below, the article and revision features are extracted directly from the text and metadata in the exported articles. The language features, including the various reading-level metrics, are calculated using the open-source Natural Language Toolkit for Python [5] and its community-contributed addendum, nltk_contrib[8]. For the creator features, we used the Wikimedia API to access the necessary information about the article creators. Finally, the topic features were collected using various external APIs; in particular, pageview statistics were obtained from an independent Wikipedia pageview tool[9].

6. FEATURES

We extract 41 features from each page in our dataset. These are divided into five groups. Selected features are discussed below; a full list can be found in Appendix A. It is important to note that, in theory, it is the encyclopedic quality of the article *topic* that is being evaluated, not the quality of the article itself. We have, however, found that many features not directly related to notability are quite discriminative in practice. We therefore attempt to use as many features as possible that can approximate the significance or notability of the topic, but also include a large number of other features which we have found helpful in discriminating between deleted and kept articles.

6.1 Article Creator Features

We hypothesize that users who create unencyclopedic articles tend to differ from those whose articles are kept. Indeed, in our previous work [14], we found significant differences between the two groups. Users whose articles were kept had, on average, been registered on Wikipedia for 5 times longer and had made 9 times more edits than those whose articles were deleted. Given these statistics, we extract the length of time the article's creator had been a registered user of Wikipedia, the number of edits he or she had made before creating the article in question, and the number of other articles they had created that were not later deleted (at the time of our analysis). We also include as a feature whether or not the creator has a user page, a Wikipedia-specific personal homepage common to legitimate editors. Finally, we determine whether the creator's account was later blocked, a disciplinary action commonly taken when a user is guilty of 'disruptive behavior [39],' such as vandalism. This feature is by definition a posteriori information, and is only of use when some time has passed since the creation of the article. (As is the case in one of our use cases: finding older articles for improvement or deletion.)

6.2 Topic Features

Topic features attempt to measure the notability or significance of the article's topic, independent of the article itself. We retrieve the number of results for each article's name from a major search engine, as well as the number of pageviews the article received over its lifetime[10]. (These have been used before as proxies for notability; see [19].) We also include the number of other web pages that link to this one – both internal Wikipedia links and all links to the page from any page on the web.

6.3 Revision Features

Each article we downloaded was accompanied by its entire revision history, which includes various metadata about each revision (edit) ever made to the page. We have found [14] that unencyclopedic articles that were Speedy Deleted tend to have fewer revisions than notable ones, and tend to be written by smaller groups of editors. We include the following features: number of revisions, number of revisions made by anonymous editors, number of revisions made by logged-in users, number of

[7] http://en.wikipedia.org/wiki/Special:Export

[8] https://code.google.com/p/nltk/downloads/detail?name=nltk_ contrib-2.0b5.zip

[9] http://stats.grok.se

[10] Unfortunately, the MediaWiki platform miscounted a large number of pageviews in December 2011, so accurate pageview statistics are only available for 2012 data.

unique editors, and whether any one editor, or specifically the article creator, has written more than 50% of the article's content. Many of the articles in our dataset appear to be autobiographical. This is frowned on by Wikipedia [34, 35], and these pages are often about non-notable people. To find some of these pages, we use the string similarity between the article title and the creator's username. About 15% of the deleted articles in the Original set had high similarity (>0.6).

6.4 Article Features

These features were extracted from the article text. They include the number of categories, images, and links to other Wikipedia pages contained in the article. These features are often added by hand, and a large number of them may indicate a better-quality article. We also extract the number of references. The basic notability standard on Wikipedia is 'multiple references' [37]; while there is no easy way to automatically determine whether an article's references are relevant and reliable, we have previously found [14] that kept articles tend to have more references than deleted ones.

6.5 Language Features

Among these are the number of nouns, verbs, adjectives, and adverbs in the article text, normalized by the length of the article. We also calculate the Flesch-Kincaid (FK) reading level score, among several others (see Appendix A) and the FK reading ease score [18]. These metrics give a rough picture of the language quality of the article. These features are quite expensive to extract. A real-time system which depends on rapid throughput, therefore, may choose not to use these features.

7. EVALUATION

To run our experiments, we use the Weka machine-learning software suite [17]. We split both the original and new datasets into training and test sets, using a 70-30 split. Model selection and parameter tuning were performed using 10-fold cross-validation over the entire training set. For the two smallest datasets, the Old set and the AfD set, we use the entire dataset for both training and testing, using 10-fold cross-validation. We use Random Forests [6] (of 40 trees) for classification because of their high accuracy and low overhead. They have also been used successfully for two similar tasks, vandalism detection [21] and flaw detection [4]. Feature selection was done using clustering and cross-validation and resulted in the elimination of several features that were decreasing classification accuracy. (See Appendix A.)

To the best of our knowledge, there is no comparable work in this area to serve as a baseline. We therefore experiment on each deletion type (Speedy, PROD, and AfD) separately and compare the results. For Speedy Deletion, where we achieved the best results, we also compare the results obtained using various combinations of features as an approximation of a baseline. For all article types, we report the precision and recall for the deleted class. In the case of deletion, precision is much more important than recall. Our goal is to assist humans in finding articles for deletion, not replace them; there will still be human editors who can find any unencyclopedic articles missed by our system. Marking a good article as deletion-worthy, however, is something we strongly want to avoid. We therefore attempt to maximize precision over recall. We concentrate on precision and recall for the deleted class since we are much more concerned about articles wrongly classified as deletion-worthy than about those wrongly classified as encyclopedic. Table 2 summarizes our results.

Table 2: Results for each deletion method, given for the deleted class.

	Precision	Recall
AfD	96%	33%
PROD	98%	71%
Speedy	98.6%	97.5%

7.1 Articles for Deletion (AfD)

As expected, detecting potential AfDs was the most difficult problem. Since these articles are of borderline encyclopedic status and require discussion and deliberation before being deleted, they are difficult to distinguish from similar articles that were also borderline but were kept. Despite this, we are still able to achieve high precision on this problem. We achieve precision of 96% at a recall of 33%. These levels are good enough to be used in an autonomous system, and definitely one like our proposed one, which utilizes human oversight. For comparison, the most prolific vandalism-removal bot on Wikipedia, which operates independently of human scrutiny, has precision set to 99% and recall of 40% [8].

7.2 Proposed Deletion (PROD)

Given that PRODs by definition are not bad enough to meet one of the Criteria for Speedy Deletion, but are unquestionably unencyclopedic, we expect that classifying PRODs is a more difficult problem than Speedies, but easier than detecting potential AfDs. This hypothesis is borne out by our experiments; we achieve precision of 98% at a recall of 71% on the PROD set, lower than that achieved on the Speedy sets, but better than our results for AfDs.

7.3 Speedy Deletion

Precision and recall numbers for the Speedy Deletion datasets were similar, and quite high, so for simplicity we report a single measure, accuracy. This is simply the percentage of articles classified correctly. Table 3 gives detailed results for the Speedy Deletion experiments – the Original, Old, and New datasets.

7.3.1 All Features

We first test the Original, time-biased dataset using all features. As expected, accuracy is very high, approximately 97%, probably due to the confounding factor of the very different ages of the articles in the two sets (see Section 5.4.5). Using all features for classifying the set of Old pages yields a more modest, but also more realistic, improvement over the baseline. Interestingly, removing all features that might be biased by the age of the page (13 features) decreased accuracy by just a few percentage points, due to the contributions of the creator and language features (see below). Finally, we test the New dataset, achieving accuracy of 95%. These latter experiments show that the good results hold true even when the effect of the age of the article is accounted for.

7.3.2 Creator Features

The group of features related to the article's creator produced strong individual results. In fact, one of them, the creator's previous edit count, was so discriminative that it alone achieved a 44% improvement over baseline on the Original dataset, and 23% on the Old set. Together, the creator features yielded 91.5% accuracy on the Original dataset, 85.5% accuracy on the Old pages set, and 92.8% on the New set. This is noteworthy because none of these features rely on the age of the article, and therefore, the accuracy on the Original dataset can be taken at face value.

Table 3: Speedy Deletion: Accuracy of classifier on each dataset, for several feature combinations.

	Original	Old	New
Baseline (Zero-R)	**63.42%**	**75%**	**72.7%**
All features	**97.57**	**92.6**	**95.21**
No language features	97.22	91.31	95.55
Language Features	96.18	93.8	78.0
Creator Features	91.49	85.47	92.8
Revision Features	95.04	82.1	83.0
Article Features	90.88	82.1	85.39
Non time-bound	95.79	N/A*	N/A*
Bag of Words (SVM)	96.55		
2011 training, 2012 test	96.4		

*Already normalized by time.

7.3.3 Language Features

The language feature group had the most variation between different datasets. While the language features alone were extremely discriminative for the Original and Old datasets, even achieving better accuracy on the Old dataset than all the features combined, they were less discriminative on the New dataset than any other feature group, even those that performed poorly on the other datasets. The part of speech features accounted for most of the accuracy on the Original and New sets. We previously found [14] that deleted articles had, on average, twice as many verbs and adverbs as kept ones, and more adjectives. We hypothesize that this may capture a fundamental difference in the way articles are written – articles deleted for lack of significance tend to use more active (verbs), descriptive (adjectives) language in discussing the activities and/or importance of their subjects, while encyclopedic articles focus on the entity itself (nouns). Given the poor performance of the language features on the New dataset, though, we are reluctant to accept any explanation without further inquiry.

The language features were by far the most expensive to extract. We therefore measure the performance of our system with and without the language features. We find that the language features were not particularly helpful, increasing accuracy by less than 1% on the Original dataset and just 1% on the Old dataset; removing the language features actually increased accuracy on the New dataset. (Possibly for the same reason that these features alone were less discriminative on the New dataset than on the others.) Since the language features performed so well on their own, this lack of significant improvement is probably due to the highly discriminative nature of some of the other features.

7.3.4 Discussion

We find that the Old dataset achieved less improvement over the baseline, and lower absolute accuracy, than the Original set. This was true even for those features that are not influenced by the age of the page, such as the creator features. We believe this resulted from the nature of the deleted pages in the set. Since the vast majority of Speedy Deleted articles are deleted within a week of their creation [33], the few that are deleted later usually have a reason for surviving so long, and are often of better quality. This makes it more difficult to differentiate between them and ordinary encyclopedic articles, and accounts for the lower accuracy on the Old dataset. The large (23%) improvement over baseline achieved even on this dataset, however, shows that our system can successfully identify unencyclopedic articles even when they have remained on Wikipedia for a long time. It can therefore be used to detect older articles for improvement or deletion, as mentioned in our list of goals. The results from the New dataset were also very interesting. Since this was the most unbiased dataset, given that the articles were of similar age, the success of our classifier on this set is the most noteworthy. In addition, while in the Original and Old datasets, there was at least one feature group that alone performed nearly as well as, or better than, the entire classifier, the performance of all the features in the New dataset was much better than any subset of features. Since we consider the New dataset to be the most unbiased and representative, this is an encouraging sign for the validity of our model.

7.3.5 Text

We also evaluated the performance of a classifier trained only on the article text. In this case only, we used a Support Vector Machine [9], because it performs very well on high-dimensional data. We used the top 1500 most frequent words in the dataset as a Bag of Words to train and test the SVM. Using the words alone, we achieved 96.6% accuracy on the Original dataset. This provides further support to the idea that there are significant inherent differences between kept and deleted pages.

7.3.6 Robustness

Since the articles in each dataset were collected over a short period of time, it is possible that a model built from them would not generalize well to future data. We train a classifier using the data from the Original dataset and test it on a set of articles deleted in November 2012 (a year after the original data collection). The classifier achieves >96% accuracy on the new data. These results suggest that the characteristics of deleted articles that we have found are not artifacts of our data and do in fact hold true over a longer period of time.

8. LIMITATIONS AND CONCERNS

While our model has been shown to have high precision (and in most cases high recall as well) on several different datasets, our ultimate goal, its deployment on Wikipedia for some or all of the functions listed in Section 4, raises several concerns. The most obvious limitation of our model is that it attempts to model a somewhat subjective decision (whether or not an article is notable) using objective criteria (features of the article and its metadata). In this limitation, though, it is no different than many other models which attempt to fit objective factors to an often subjective, complex phenomenon, and the same caveats apply as do to all such modeling. However, we also used articles that *were* deleted as a gold standard to represent articles that *should be* deleted – that is, accepting without question that the decisions made by article deleters are correct, which may not be true. Since our gold-standard set is actually very subject to human judgment, it is possible that, rather than our features accurately modeling unencyclopedic articles, articles are being deleted *because* they contain or lack these features, whether or not they are actually encyclopedic. (For instance, having a large number of verbs was a strong predictor of deletion; it is possible that unencyclopedic articles tend to contain many verbs, but also possible that editors tend to delete articles with many verbs without checking well enough for the significance of the article topic.) If this is so, using our model to classify articles as deletion-worthy or not will just compound this problem. In a similar vein, if the model does not accurately model topic significance, when used as a self-evaluation tool, article creators can work out how to “game the system” by changing minor properties of the article to make it fit the (erroneous) model better.

The only way, then to determine the accuracy of the model would be a manual examination of the deleted articles in our dataset to determine whether or not they were correctly deleted, as well as an examination of the output of our classifier on a large set of articles. If the articles in the dataset are found to have been legitimately deleted, this would validate our use of them as a gold standard, and, by extension, our model. We plan to carry out this analysis in the future, perhaps as part of preliminary testing of our proposed tool.

Another potential, and serious, concern relates to part of the motivation for our work, the issue of newcomer retention. An increasing body of research has shown that the use of automated tools has a negative impact on newcomer retention [13, 16]. Given the number of newcomers who are lost due to overly-aggressive deletion of their articles using automated tools [16], it would seem that the last thing Wikipedia needs is another automated tool assisting in the deletion process. However, our system is different than the ones implicated in scaring off newbies. The main issue there seems to be that the *interaction* between new users and more experienced ones has been reduced because of these tools; tool-generated communication also tends to be harsher than human-to-human interaction [13]. The tools studied were either completely automated bots or assisted editing tools, which allow a user to make a single decision (delete, revert, etc.) and then automatically take care of the rest of the ‘housekeeping’ involved, including notifying the offending editor that his content has been removed. When the receiving editor attempts to discuss the issue, he or she is generally ignored [13], which breeds resentment and often leads new editors to abandon the site [11]. The key difference between these tools and our proposed one is that they take human input at the *beginning* of the process and then carry out the rest of the task – including such crucial human aspects as inter-editor communication – automatically. Our proposed system, on the other hand, will deal with the first part of the decision process – preliminary information gathering and analysis – and then have human editors take care of the *end* of the task, including the communication aspects. We envision our tool as a decision-support system rather than an assisted editing tool. This will hopefully give users all the benefits of decision support (faster, more informed decision-making) without increasing aggressiveness towards new editors – in fact, we argue that it may even decrease it (see Section 4).

Unlike the creators of previous tools, we also have the benefit of knowledge about the impact of automated tools on editor retention. Using this knowledge, we can carefully design our tool(s) to guard against many of the problems mentioned above (for instance, by not giving the option of automated notifications or warnings) and carefully deploy it - perhaps only to a small, trusted subset of experienced editors - while continuously evaluating its impact on the Wikipedia community and making any necessary changes. This will also help mitigate the possibility that users will simply accept the decisions of the tool without doing the necessary checks that it is indeed correct. Given the delicate nature of AfD decisions in particular, and the widespread (and possibly correct) perception that they require human judgment, we may decide to deploy our system only for the detection of Speedy Deletion candidates. In that case, the AfD and PROD experiments reported in this paper would serve as a proof of concept that our model is indeed robust.

Ultimately, we believe that the benefits of a carefully designed, cautiously deployed set of tools might outweigh the possible negatives. In the words of the Wikimedia Foundation itself, “Technical changes can't easily force a patroller to spend a certain amount of time working on each page, and even if they could it would be unfair. Instead, our focus is on ensuring that if patrollers are going to patrol things quickly, they can patrol things quickly and properly. There shouldn't need to be a tradeoff between speed and quality” [32]. Each of our suggested uses has the potential to benefit many Wikipedians, and Wikipedia itself, in a significant way if successful. However, the use of automated tools on collaborative systems like Wikipedia is a complex and ever-changing topic which merits continuous dialogue.

9. CONCLUSIONS AND FUTURE WORK

We find that we can predict which Wikipedia articles will be deleted with a high degree of precision, even after accounting for the potential confounding effect of the age of the article. We achieve accuracy of up to 97%, and precision of 98% with recall of 97%, in the case of Speedy Deletions, the most clearly unencyclopedic articles. Our accuracy is lower for PRODs and AfDs, which are less egregiously bad, but we still achieve precision of 96% at recall of 33% on the most difficult problem, AfDs. This performance is good enough for our model to be used for several different uses – helping editors find articles to improve, assisting article creators in bringing their articles to Wikipedia standards, and as a decision-support tool to assist in determining whether or not articles should be deleted. We plan to implement our system as a set of add-on Wikipedia tools, keeping in mind the caveats discussed in the previous section.

Although our results using the current features were very good, we would like to experiment with various groups of features, including some we have not yet implemented, such as the time of article creation and creator location, to find the feature set that optimizes both performance and throughput. We are particularly interested in using topic modeling to improve our results, since the encyclopedic status of an article depends mostly on its topic.

Over the course of the investigation described in this paper, we found many deletion candidates that seemed to have been misclassified (though generally still deserving deletion, but for a reason other than the stated one). We would like to make a comprehensive review of the various forms of deletion in Wikipedia to determine whether or not they are being applied correctly. As mentioned above, this would help us determine how accurate our gold-standard data, and therefore our model, is. We would also like to analyze the AfD dataset to see if there are quantifiable differences between kept and deleted AfDs.

8. ACKNOWLEDGEMENTS

Many thanks to the anonymous reviewers for going beyond their stated duty to improve this paper. Research supported by GAANN Grant P200A090157 from the US Department of Education.

10. REFERENCES

1. Adler, B., de Alfaro, L., Pye, I. Detecting Wikipedia Vandalism using WikiTrust. In Braschler, M., Harman, D., eds. *Notebook Papers of CLEF 2010 LABs and Workshops*. (2010)

2. Adler, B.T., et. al. Wikipedia Vandalism Detection: Combining Natural Language, Metadata, and Reputation Features. *Proc. CICLing 2011,* 277-288.

3. Alexa.com Site Rankings. http://www.alexa.com/siteinfo/wikipedia.org.

4. Anderka, M., Stein, B., and Lipka, N. Predicting Quality Flaws in User-generated Content: The Case of Wikipedia. *Proc. SIGIR 2012.*

5. Bird, S., Klein, E., and Loper, E. *Natural Language Processing with Python.* O'Reilly Media, 2009.

6. Breiman, L. Random forests. *Machine learning*, *45*(1), (2001). 5-32.

7. Burke, M., and Kraut, K. Mopping up: Modeling Wikipedia promotion decisions. *Proc. CSCW 2008.*

8. ClueBot NG. http://en.wikipedia.org/wiki/User:ClueBot_NG

9. Cortes, C. and Vapnik, V. Support vector networks. Machine Learning, 20:3, 1995, pp. 273-297.

10. Editor Trends Study. http://strategy.wikimedia.org/wiki/Editor_Trends_Stud y

11. Ford, H., and Geiger, R.S. "Writing up rather than writing down": Becoming Wikipedia literate. *Proc. WikiSym 2012*.

12. Geiger, R. S. and Ford, H. Participation in Wikipedia's article deletion processes. *Proc. WikiSym 2011.*

13. Geiger, RS., et. al. Defense mechanism or socialization tactic? Improving Wikipedia's notifications to rejected contributors. *Proc. ICWSM 2012*.

14. Gelley, B. Investigating deletion in Wikipedia. 2013. *In manuscript.* Available at http://arxiv.org/abs/1305.5267.

15. Halfaker, A, Kittur, A., and Riedl, J. Don't bite the newbies: how reverts affect the quantity and quality of Wikipedia work. *Proc. Wikisym 2011.*

16. Halfaker, A., et. al. The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline. *American Behavioral Scientist,* 2013, 57:664. Published pre-print online Dec. 28, 2012.

17. Hall, M., et. al. The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1, 2009.

18. Kincaid, J. P., et. al. (1975). Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel (No. RBR-8-75). *Naval technical training command Millington, Tenn research branch.*

19. Lam, S.K., and Riedl, J. Is Wikipedia growing a longer tail? *Proc. GROUP 2009*, pages 105–114.

20. Lam, S. K.., Karim, J., and Riedl, J. The effects of group composition on decision quality in a social production community. *Proc. GROUP 2010*, pages 55–64.

21. Mola-Velasco, S.M. Wikipedia vandalism detection through machine learning: Feature review and new proposals. In Braschler, M., Harman, D., eds.: *Notebook Papers of CLEF 2010 Labs and Workshops. (2010)*

22. Potthast, M., Stein, B. and Gerling, R. Automatic vandalism detection in Wikipedia. *ProC. ECIR'08*.

23. Potthast, M., Stein, B., and Holfeld, T. Overview of the 1st International Competition on Wikipedia Vandalism Detection. In Martin Braschler and Donna Harman, editors, *Notebook Papers of CLEF '10 Labs and Workshops*, 2010.

24. Schneider, J., Passant, A., and Decker, S. Deletion discussions in Wikipedia: Decision factors and outcomes. *Proc. WikiSym 2012*.

25. Suh, B., Convertino, G., Chi, E. H., and Pirolli, P. The singularity is not near: slowing growth of Wikipedia. *Proc. WikiSym 2009.*

26. Taraborelli, D. and Ciampaglia, G. L. Beyond notability. Collective deliberation on content inclusion in Wikipedia. *Proc. SASOW 2010*.

27. User:Mr.Z-man. New user study. http://en.wikipedia.org/wiki/User:Mr.Z-man/newusers.

28. Wang, W.Y. and McKeown, K. "Got You!": Automatic vandalism detection in Wikipedia with web-based shallow syntactic-semantic modeling. *Proc. Coling 2010*, pages 1146–1154.

29. West, A.G., Kannan, S., Lee, I.: Detecting Wikipedia vandalism via spatio-temporal analysis of revision metadata. *Proc*. *EUROSEC 2010.*

30. West, A. G. and Lee, I. What Wikipedia deletes: Characterizing dangerous collaborative content. *Proc. WikiSym 2011*, pages 25–28.

31. Wikimedia Foundation, Article Creation Workflow. http://www.mediawiki.org/wiki/Article_creation_workflow Supporting_evidence

32. Wikimedia Foundation. New Page Patrol Survey. http://meta.wikimedia.org/wiki/Research:New_Page_Patrol_ survey/WMF_report

33. Wikimedia Foundation Research; The Speed of Speedy Deletions. http://meta.wikimedia.org/wiki/Research:Th e_Speed_of_Speedy_Deletions.

34. Wikipedia:Autobiography. http://en.wikipedia.org/wiki/Wikipedia:Autobiography#Creating_an_article_about_yourself

35. Wikipedia, Conflict of Interest Policy. http://en.wikipedia.org/wiki/Wikipedia:Conflict_of_interest.

36. Wikipedia:Criteria for Speedy Deletion. http://en.wikipedia.org/wiki/Wikipedia:Criteria_for_speedy_ deletion.

37. Wikipedia:Notability. http://en.wikipedia.org/wiki/Wikipedia:Notability

38. Wikipedia New Page Patrol Talk Page, Archive 2. http://en.wikipedia.org/wiki/Wikipedia_talk:New_pages_patrol/Archive_2#Speedy_speedy_tagging

39. Wikipedia policy on Blocking. http://en.wikipedia.org/wiki/Wikipedia:Blocking_policy

40. Wikipedia Size Comparisons. http://en.wikipedia.org/wiki/Wikipedia:Size_comparisons

41. Williams, S. Summaries of Wikipedia deletion discussions: a shallow semantic approach. Senior Thesis, 2006, University of Colorado at Boulder. Advisor James H. Martin. Available from Citeseer only.

APPENDIX A – FEATURES USED FOR CLASSIFICATION

Feature Name	**Description**
Creator Features	
Creator Name*	Username of article creator
Creator Days on Site	How long has article creator had a WP account?
Creator Num Edits	How many edits has the article creator made sitewide?
Creator Status	Creator's account status (open or blocked)
Num other pages	Number of other kept pages created by article creator
Userpage	Does article creator have a userpage?
Page Features	
Title*	Article Title
Createdate*	Date article was created
File Size[!]	Size of the entire file, including all revisions
Has Talk Page[!]	Does the article have an accompanying talk page
Topic Features	
Num links to here	Number of other Wikipedia articles linking to the article
Num links in from Web	Number of external web pages linking to article
Pageviews**	Number of visits page has had
Num Hits	Number of search engine results for the page title
Article Features	
Num categories	Number of WP categories the article belongs to
Num images	Number of images in the article
Num references	Number of references in the article
Num sections	Number of sections in the article
Num out Wikilinks	Number of links in the article to other WP pages
Infobox*	What kind of infobox does the article contain?
Total Size in Bytes	Length, in bytes, of the final version of the article
Revision Features	
Num Revisions	Number of revisions to the article
Num registered edits	Number of edits made by registered users
Num anonymous edits	Number of edits made by anonymous users
Num unique Editors	Number of unique users who edited the article
Time to Delete	Number of days between article creation and proposal for deletion
More than half anon	Boolean; are more than half of the edits made by anon users?
Has main editor	Boolean; has one user created > 50% of the content?
Creator is main editor	Boolean; is the article creator the main editor?
Likelihood Autobio	String similarity between article title and creator's username

Text*	Bag of all words in the final version of the article
Language Features	
Normalized noun count	# of nouns in article, normalized by article length
Normalized verb count	# of verbs in article, normalized by article length
Normalized adjective count	# of adjectives in article, normalized by article length
Normalized adverb count	# of adverbs in article, normalized by article length
FK reading level	Flesch-Kincaid reading level of the article
SMOG reading level	SMOG reading level index
Cl level	Coleman-Liau reading level index
Level avg	Average of 5 reading level indexes
FK reading ease	Flesch-Kincaid reading ease measure of the article

* Not used for classification because these features are extremely sparse.

** Only used in New dataset because of a MediaWiki software error that miscounted pageviews in December 2011.

! Not used in the Original dataset classification because feature selection found that they reduced accuracy.