Detecting Vandalism in Open Source A Case Study: Wikipedia

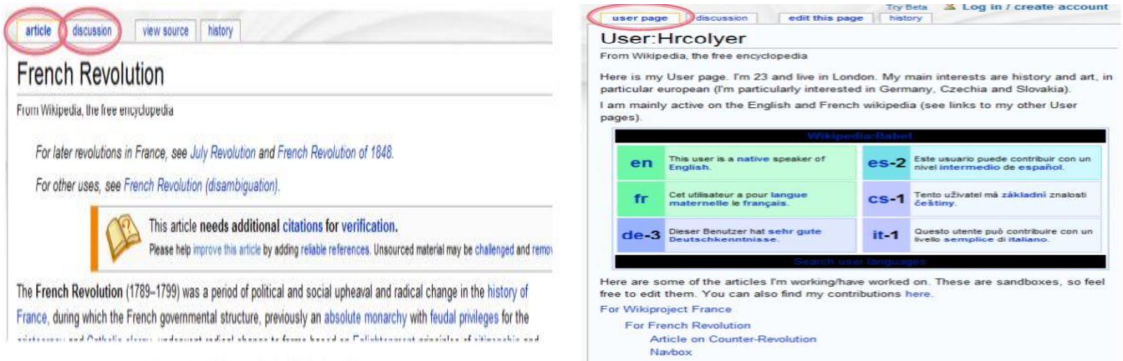
Abu Saleh Md Noman
School of Informatics and Computing
Indiana University Bloomington
amdnoman@indiana.edu

Abstract- **Modern online communities are experiencing foul edits, spam, and vandalism. A lot of online discussion communities nowadays offer users to hide their identity and interact. Such flexibility is understandable, however, they engender threats to the reputation and reliability in collective goods. In this study I am interested in studying one such open source online community: Wikipedia. Wikipedia is the largest open and free online encyclopaedia that is accessible to all. I am especially interested in finding the type of activity that the anonymous users are doing on Wikipedia articles and can we predict vandalism in Wiki articles using NLP based Machine Learning algorithms. Since not a lot of previous work addressed these issues it is important to study the aforementioned issues to build an innate understanding of recent ongoing vandalism of Wikipedia pages and editors leaving Wikipedia. The size of Wikipedia is vast (approximately 28,085,962 users, 5,135,196 English articles and 39,169,255 wiki pages) and that makes the study of Wikipedia more interesting. The study reveals ~90% of the vandalism or foul edits are done by unregistered users in Wikipedia thanks to the free and open nature of it. I compiled a large number of vandalism edits in a corpus, which allows for the comparison of existing and new detection approaches. Using logistic regression I achieve 83% precision at 77% recall with my model. Compared to the rule-based methods that are currently applied in Wikipedia, this approach increases the F-Measure performance by 49% while being faster at the same time**

I. INTRODUCTION

Wikipedia is the largest multilingual, web-based, free-content encyclopedia that is available in the internet. This project is supported by Wikimedia foundation and it is completely open source that means it is editable by anyone. Since its birth in 2001, it has experienced a massive growth in size and now it is invoking 374 million unique visitors monthly as of September 2015 [24]. About 70,000 active contributors are there working on more than 38 million articles in almost 292 languages. As of today there are 5,135,190 articles in English (compared to 3.9 million articles to nearest competitor Baike.com, Chinese wiki). Everyday there are tens of thousands of edits and visits to different Wikipedia pages help it grow and the gigantic size and diversity in Wikipedia data makes it particularly intriguing for researchers interested in mining patterns [1] and trends in the data. People from all backgrounds, cultures and ages can view and edit existing pages that makes the collaboration regardless of their qualification, it is the content not the quantity that matters. The reliability of the edits are reviewed by bots and experienced Wikipedians ensuring the quality of content that pertains.

However, the credence of such contribution are often questionable because of this open nature. While articles or pages evolve over the period of time and the collaboration making them more comprehensive, they are vulnerable to misinformation, errors, vandalism, foul edits, hacks etc.. Every Wikipedia article has an ‘Edit’ and ‘View History’ button (Figure 1(a)) that allows user to make necessary changes, summarize them and publish. They can make minor edits (spelling correction, grammatical changes) that require no further attention or they can make major edits which require further scrutiny by the community. Some pages are semi-protected indicated by the lock sign but still can be edited by placing an edit request. This makes these articles vulnerable to attack by foul editors and we need to devise a way to detect such vandalism algorithmically since manual moderation is not feasible.



French Revolution

From Wikipedia, the free encyclopedia

For later revolutions in France, see July Revolution and French Revolution of 1848.

For other uses, see French Revolution (disambiguation).

This article **needs additional** citations for verification.
Please help improve this article by adding reliable references. Unsourced material may be challenged and remo

The **French Revolution** (1789–1799) was a period of political and social upheaval and radical change in the history of France, during which the French governmental structure, previously an absolute monarchy with feudal privileges for the

User:Hrcolyer

From Wikipedia, the free encyclopedia

Here is my User page. I'm 23 and live in London. My main interests are history and art, in particular european (I'm particularly interested in Germany, Czechia and Slovakia).

I am mainly active on the English and French wikipedia (see links to my other User pages).

Wikipedia:Babel			
en	This user is a native speaker of English.	es-2	Este usuario puede contribuir con un nivel intermedio de español.
fr	Cet utilisateur a pour langue maternelle le français.	cs-1	Tento uživatel má základní znalosti češtiny.
de-3	Dieser Benutzer hat sehr gute Deutschkenntnisse.	it-1	Questo utente può contribuire con un livello semplice di italiano.
Search user languages			

Here are some of the articles I'm working/have worked on. These are sandboxes, so feel free to edit them. You can also find my contributions here.

For Wikiproject France
For French Revolution
Article on Counter-Revolution
Navbox

Figure 1. (a) A typical Wiki page (b) A typical user page (image taken from Wikipedia)

Reviewing and moderating such a large scale of revisions require extensive time and human effort. In this paper, I contribute a new machine learning-based approach for vandalism detection in Wikidata, and so I wish to minimize the effort of human moderators of detecting vandals manually and painstakingly. I analyzed features suitable for Wikidata by considering content and contextual information of edits. I applied logistic regression based approach by 10-fold cross validation to discriminate vandalism and good edits. Wikidata Vandalism Corpus WDVC-2015 [25] comprising of 24 million revisions which is readily available was used.

Research challenges. This work addressed various challenges for making the study possible. First, the dataset of Wikipedia dump is of titanic scale. The English wiki dump produces terabytes of edit history which was cumbersome to handle. Second, the results of analysis are often misleading and it is difficult to determine a comprehensive metric for the study. Third, due to the careful edits by vandals off late, the classification task is ever more challenging

Contribution.

In this paper I develop basis for an automatic vandalism detection in Wikipedia: (i) I define vandalism detection as a classification task, (ii) discuss vandalism is defined by humans (iii) extract features out of them (v) Finally, the evaluation and discussion of findings based on this corpus.

II. RELATED WORK

Adler et. al [18] presented the results of an effort to integrate three of the leading approaches to Wikipedia vandalism detection: a spatio-temporal analysis of metadata (STiki), a reputation-based system (WikiTrust), and natural language processing features. They examined in detail the contribution of the three approaches, both for the task of discovering fresh vandalism, and for the task of locating vandalism in the complete set of Wikipedia revisions. Potthas [19] presented results of a new approach to detect destructive article revisions, so-called vandalism, in Wikipedia. They discussed the characteristics of vandalism as humans recognize it and develop features to render vandalism detection as a machine learning task.

Adler et. al [20] presented using the full set of features computed by WikiTrust, they have been able to construct classifiers that identify vandalism with a recall of 83.5%, a precision of 48.5%, and a false positive rate of 8%, for an area under the ROC curve of 93.4. Using these classifiers, they have implemented a simple Web API that provides the vandalism estimate for every revision of the English Wikipedia. A statistical language model, constructing distributions of words from the revision history of Wikipedia articles was presented in [21]. As vandalism often involves the use of unexpected words to draw attention, the fitness of a new edit when compared with language models built from previous versions may well indicate that an edit is a vandalism instance. Neis et al. [26] analyzed how rule based approach can be used to find foul edits in OpenStreetMap. The score they used marks (1) edit reputation, and (2) user reputation, respectively. While English Wiki was well studied, West and Lee [27] propose a set of language-independent features, and Tran and Christen [28] attempt transfer learning between languages. This is particularly useful when applied on minority language editions of Wikipedia.

III. Data and methods

For the purpose of user behavior & types of edits and vandalism study I collected 2 different dataset:

i.

The data is collected from the huge data dumps those are already available in Wikimedia sites (https://dumps.wikimedia.org/enwiki/). The dumps for different language versions of Wikipedia are kept separate. It is important to note that the data are available in XML format (some previous version have SQL and HTML dumps too but are out of date) and needed to be transformed into more readable format. So I had to execute a Java tool that is already available called *mwdumper* that converts the XML to SQL. Some other tools such as mwdump.py, ImportDump.php, xml2sql (https://meta.wikimedia.org/wiki/Data_dumps/Tools_for_importing) are also available. But considering the huge size of English wiki dump, *mwdumper* is the best solution that generates the script without getting crashed. After initial analysis, data was loaded into a MySQL database and relevant results were extracted. Some recent dumps are also available in JSON format too. Loading JSON files in MongoDB and analyzing with pyMongo can greatly reduce the processing time.

There are three main reasons behind why Wikipedia dataset was chosen for this particular study: First, wiki dataset is publicly available in chunks and it is ideal for comprehensive longitudinal study. Second, Wikipedia data offers the diversity unlike anything else. All data are representative of sample drawn from diverse population. Third, the dataset is apparently harmless since it contains no personal information about user. For the sake of the study and brevity of analysis, the dataset used was ‘enwikisource-20160305-pages-meta-history.xml’ (https://dumps.wikimedia.org/enwiki/) combined with another dump of Bengali Wikipedia ‘bnwikibooks-20160407-pages-meta-history.xml’ from another timeframe. This was done to introduce more diversity in the data. The whole dataset was divided into 50 different tables under a predefined schema. The three main tables are given by:

- *user* – gives the names and total number of edits for each used in the dataset;
- *page* – provides information about the Wikipedia pages in the dataset;
- *revision* – saves the revision of each page with user id, comments and timestamp.

The remaining tables describe the page/user categories and their relations. Table pagecategory gives information about page categories. Similarly, table category provides information about user categories. Table 1 summarizes the dataset:

Dataset	# of revisions	#of Registerd Users	#of Anonymous Users
20160305	158112	2790	5741
20160407	148399	2513	5921

Table 1: Dataset Summary

ii.

Vandalism Evaluation Data:

The Wikidata Vandalism Corpus WDVC-2015 [25] is mostly popular and widely used publicly available database on vandalism. It contains all of about 24 million revisions that were manually created between October 2012 (when Wikidata went operational) and October 2014, disregarding revisions created automatically by bots. 103,205 revisions were labeled as vandalism if they were reverted using a tool dedicated to revert vandalism [29]. About 18% of users have vandalized at least once and 1% of items were targeted at least once. 86% of the revisions labeled as vandalism are true vandals and only about 1% of the revisions labeled non-vandalism are reverted manually. After analyzing the corpus it was seen that 24 million revisions were created by 299,000 unique users editing about 7 million different items in about 14 million work sessions.

The dataset was not split into test, validation or training subsets. If we take data at random that might be problematic because some later revision might be used to train the classifier that eventually classifies an earlier edit. So it should be noted that the classifier has to perform the task in a chronological order. Keeping that in mind I separated the dataset based on the time they were performed. The feature selection and parameter tuning was done only based on validation set.

The metric used for vandalism study is the percentage of posts marked as vandalism, therefore,

$$\frac{revisions\ marked\ as\ vandalism}{total\ number\ of\ revisions} * 100\%$$

Let $E = \{e_1, \ldots, e_n\}$ denote a set of edits, in chronological order such that, $e = (d_t, d_{t+1})$. Let $F = \{f_1, \ldots, f_p\}$ is a set of features which mark vandalism such that each feature f_i is a function that goes from edit to R $(real\ numbers)$, $f_i : E \rightarrow R$. Suppose in the corpus E we have a substantial representation of vandal edits compared to non-vandals. A classifier that outputs 0(non-vandal) or 1 (vandal) is trained with examples from E. If we have the featureset and the classifier c; it is straightforward to get the outcome $c(e)$ whether or not it is a vandalism.

Featureset Generation:

Based on the study [19], I used the features F which quantify the characteristics of vandalism in Wikipedia:

Feature *f*	Description
Char distribution	deviation of the edit's character distribution from the expectation
Char sequence	longest consecutive sequence of the same character in an edit
Compressibility	compression rate of an edit's text
Upper case ratio	ratio of upper case letters to all letters of an edit's text
Term ratio	average relative frequency of an edit's words in the new revision
Longest word	length of the longest word
Pronoun frequency	number of pronouns relative to the number of an edit's words
Pronoun impact	percentage by which an edit's pronouns increase the number of pronouns in the new revision
Vulgarism frequency	number of vulgar words relative to the number of an edit's words
Vulgarism effect	percentage by which an edit's vulgar words increase the number of vulgar words in the new revision
Size ratio	the size of the new version compared to the size of the old one
Replacement similarity	similarity of deleted text to the text inserted in exchange
Context relation	similarity of the new version to Wikipedia articles found for keywords extracted from the inserted text
Anonymity	whether an edit was submitted anonymously, or not
Comment length	the character length of the comment supplied with an edit
Edits per user	number of previously submitted edits from the same editor or IP

IV. RESULTS

This section is devoted to the findings of the analysis performed on the dataset mentioned above. The results are depicted using graphs and tables. The analysis can be grouped according to different dataset and corresponding research questions. So following subsections will try to elucidate the results:

A. *Type of Edit*

At first I analyzed the type of article both type of users are targeting. The result is depicted in Figure 2.

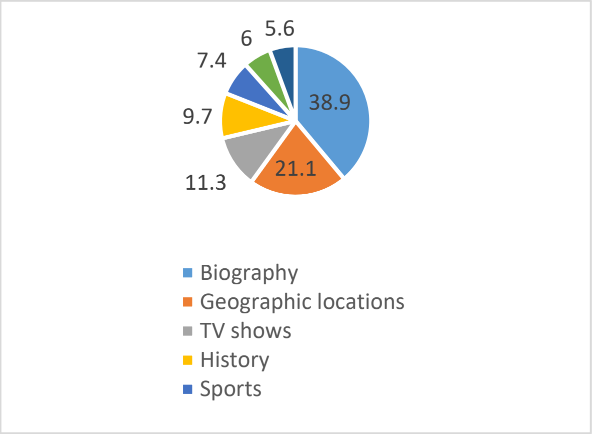


Figure 2. Type of articles

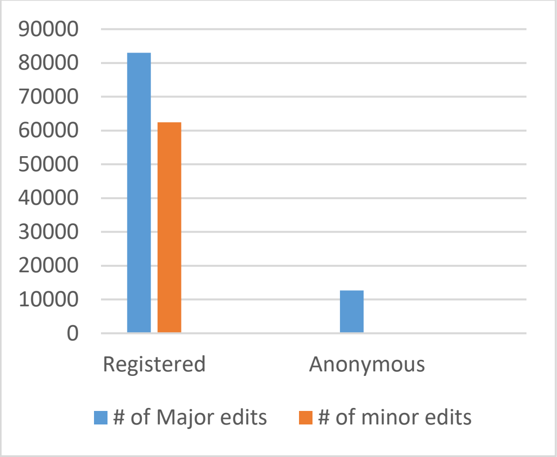


Figure 3. Type of edits

It turns out that majority of the articles targeted by registered users are related to Biography (38.9%), followed by Geographic location (21.1%), TV shows, History, Sports, Politics etc..

B. *Vandalism Study*

The direction of the research shifted towards vandalism study at this point because it was one of the most important research questions. Vandalisms are often caused by lack of knowledge, attention seeking attitude, personal grudge etc.. It is important to understand not every malicious act is considered as vandalism. Things like abuse of tags, illegitimate page creation, spam external linking, trolling etc. are considered few of different types of vandalisms. The vandalism study was performed on different dataset: for randomly selected articles and querying the API.

For randomly selected articles the manual analysis yielded following results:

TABLE III

Year	#of Articles	# of Edits	# of Vandalism	#Caused By Anonymous Users
2010	40	450	34	25
2011	50	397	29	18
2012	35	869	59	45

On an average, ~90% of times the vandalism are caused by anonymous users as predicted. However, study on user pages yielded interesting result. Out of 10 randomly generated user page, the ratio (% of vandalism done by registered to anonymous) returned was 53:47. This might be indicative of the fact that, anonymous user tend to target main article pages while registered users are main culprit for vandalisms in user pages. The analysis on *20160407* dataset yielded similar results: out of 156 commented vandalisms, 124 were done by IP users consistent with the previous finding.

C. Machine Learning Outcomes

Baselines. The baselines for our analysis are the Wikidata Abuse Filter (FILTER) [30]. The abuse filter implements some rules which might be invoked and ask for further review. The performance of the abuse filter is not well known because it is being used for quite sometime now yet the performance is not tested. The model that I used also takes into account the revisionTag feature.

Performance measures. A commonly used performance metric is area under curve of the receiver operating characteristic ($ROC_{AUC)}$ [31, 32]. Precision-Recall is a useful measure of success of prediction when the classes are very imbalanced. In information retrieval, precision is a measure of result relevancy, while recall is a measure of how many truly relevant results are returned. The precision-recall curve shows the tradeoff between precision and recall for different threshold. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. High scores for both show that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall).

- Precision= tp/(tp+fp)
- Recall = tp/(tp+fn)

One of the major issues while training a one-class classifier is the fact that sometimes there are not enough examples in one class to train. In a realistic detection scenario only 5% of all edits in a given time period are from the target class “vandalism” [33]. To solve the issue, when training, sometimes I had to randomly oversample the underrepresented class. The training examples of the sample required exhaustive analysis based on domain specific information. 10-fold cross validation was used with logistic regression classifier using the features described above. The precision and recall and ROC_{AUC} is given below for the classifier.

Feature *f*	Recall	Precision	ROC_{AUC}
Baseline	.40	.77	.575
With all features	.88	.87	.521
Char distribution	.04	.45	
Char sequence	.02	.43	
Compressibility	0	.68	
Upper case ratio	.17	.29	
Term ratio	0	.62	
Longest word	0	.3	
Pronoun frequency	.010	.55	
Pronoun impact	0	.6	
Vulgarism frequency	.25	.41	
Vulgarism effect	.24	.66	
Size ratio	.09	.90	
Replacement similarity	-	-	
Context relation	0	.23	
Anonymity	0	0	
Comment length	0	0	
Edits per user	.93	.67	

It shows the classifier outperforms baseline and rule based approach. Individual row below that show how each feature contribute to the accuracy. The high recall means less false negatives, so manual inspection of edits is not impossible. If we can apply both high precision and high recall classifier along with a confidence score, this will significantly reduce the necessity of manual review.

V. Conclusion

In this work I tried to address the automatic detection of Wikipedia vandalism by Machine Learning based approach. I argued how content and context based features are cruicial to detect and train the model. I experimented only with logistic regression which is a major drawback of this work. I would like extend this approach with parameter optimized random forest and see how it might outperform the state of the art classifiers.

For future work, I would like to analyze how Deep learning can be used to efficiently detect vandalism before the articles/edits are posted. Another good work could be giving a reasonable description of why an edit is marked as vandalism so that editors are not discouraged once their edits get reverted.

REFERENCES

[1] Maass, D. (2013). Data Mining Revision Controlled Document History Metadata for Automatic Classification.

[2] Daxenberger, J., & Gurevych, I. (2013, October). Automatically Classifying Edit Categories in Wikipedia Revisions. In*EMNLP*(pp. 578-589).

[3] Anthony, D., Smith, S. W., & Williamson, T. (2009). Reputation and reliability in collective goods the case of the online encyclopedia Wikipedia.*Rationality and Society*,*21*(3), 283-306.

[4] Lieberman, M. D., & Lin, J. (2009, May). You Are Where You Edit: Locating Wikipedia Contributors through Edit Histories. In*ICWSM*.

[5] Wierzbicki, A., Turek, P., & Nielek, R. (2010, July). Learning about team collaboration from Wikipedia edit history. In*Proceedings of the 6th International Symposium on Wikis and Open Collaboration*(p. 27). ACM.

[6] Tsikerdekis, M. (2013). The effects of perceived anonymity and anonymity states on conformity and groupthink in online communities: A Wikipedia study. *Journal of the American Society for Information Science and Technology*, *64*(5), 1001-1015.

[7] Panciera, K., Halfaker, A., & Terveen, L. (2009, May). Wikipedians are born, not made: a study of power editors on Wikipedia. In *Proceedings of the ACM 2009 international conference on Supporting group work* (pp. 51-60). ACM.

[8] Amichai-Hamburger, Y., Lamdan, N., Madiel, R., & Hayat, T. (2008). Personality characteristics of Wikipedia members. *CyberPsychology & Behavior*, *11*(6), 679-681

[9] Lucas, M. M., & Borisov, N. (2008, October). Flybynight: mitigating the privacy risks of social networking. In *Proceedings of the 7th ACM workshop on Privacy in the electronic society* (pp. 1-8). ACM.

[10] Fang, L., & LeFevre, K. (2010, April). Privacy wizards for social networking sites. In *Proceedings of the 19th international conference on World wide web* (pp. 351-360). ACM.

[11] Madden, M. (2012). Privacy management on social media sites. *Pew Internet Report*, 1-20.

[12] Narayanan, A., & Shmatikov, V. (2009, May). De-anonymizing social networks. In *Security and Privacy, 2009 30th IEEE Symposium on* (pp. 173-187). IEEE.

[13] Szongott, C., Henne, B., & von Voigt, G. (2012, June). Big data privacy issues in public social media. In *Digital Ecosystems Technologies (DEST), 2012 6th IEEE International Conference on* (pp. 1-6). IEEE.

[14] Tsikerdekis, M. (2013). The effects of perceived anonymity and anonymity states on conformity and groupthink in online communities: A Wikipedia study. *Journal of the American Society for Information Science and Technology*, *64*(5), 1001-1015.

[15] Leskovec, J., Huttenlocher, D. P., & Kleinberg, J. M. (2010, April). Governance in social media: A case study of the Wikipedia promotion process. In *ICWSM*.

[16] Rad, H. S., & Barbosa, D. (2012, August). Identifying controversial articles in Wikipedia: A comparative study. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration* (p. 7). ACM.

[17] Kittur, A., Chi, E., Pendleton, B. A., Suh, B., & Mytkowicz, T. (2007). Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *World wide web*, *1*(2), 19.

[18] Adler, B. T., De Alfaro, L., Mola-Velasco, S. M., Rosso, P., & West, A. G. (2011). Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In *Computational linguistics and intelligent text processing* (pp. 277-288). Springer Berlin Heidelberg.

[19] Potthast, M., Stein, B., & Gerling, R. (2008). Automatic vandalism detection in Wikipedia. In *Advances in Information Retrieval* (pp. 663-668). Springer Berlin Heidelberg.

[20] Adler, B., De Alfaro, L., & Pye, I. (2010). Detecting Wikipedia vandalism using wikitrust. *Notebook papers of CLEF*, *1*, 22-23.

[21] Chin, S. C., Street, W. N., Srinivasan, P., & Eichmann, D. (2010, April). Detecting Wikipedia vandalism with active learning and statistical language models. In *Proceedings of the 4th workshop on Information credibility* (pp. 3-10). ACM.

[22] Harpalani, M., Hart, M., Singh, S., Johnson, R., & Choi, Y. (2011, June). Language of vandalism: Improving Wikipedia vandalism detection via stylometric analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2* (pp. 83-88). Association for Computational Linguistics.

[23] Rizoiu, M. A., Xie, L., Caetano, T., & Cebrian, M. (2015). Evolution of Privacy Loss in Wikipedia. *arXiv preprint arXiv:1512.03523*. Computational Linguistics.

[24] https://en.Wikipedia.org/wiki/Wikipedia:About

[25] S. Heindorf, M. Potthast, B. Stein, and G. Engels. Towards Vandalism Detection in Knowledge Bases: Corpus Construction and Analysis. SIGIR 2015.
[26] P. Neis, M. Goetz, and A. Zipf. Towards Automatic Vandalism Detection in OpenStreetMap. ISPRS International Journal of Geo-Information, 2012.

[27] A. West and I. Lee. Multilingual Vandalism Detection Using Language-Independent & Ex Post Facto Evidence. CLEF Notebooks 2011.
[28] K.-N. Tran and P. Christen. Cross Language Prediction of Vandalism on Wikipedia Using Article Views and Revisions. PAKDD 2013.
[29] Wikimedia Foundation. Wikidata:Rollbackers. https://www.wikidata.org/wiki/Wikidata:Rollbackers, 2016
[30] Wikimedia Foundation. Wikidata Abuse Filter. https://www.wikidata.org/wiki/Special:AbuseFilter, 2015.
[31] J. Davis and M. Goadrich. The relationship between Precision-Recall and ROC curves. ICML 2006.
[32] H. He and E. Garcia. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21(9):1263–1284, Sept. 2009.
[33] Priedhorsky, R., Chen, J., Lam, S., Panciera, K., Terveen, L., Riedl, J.: Creating, Destroying, and Restoring Value in Wikipedia. In: Group 2007 (2007)
[34] Heindorf, S., Potthast, M., Stein, B., & Engels, G. (2016, October). Vandalism detection in Wikidata. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (pp. 327-336). ACM.