Cross Language Prediction of Vandalism on Wikipedia Using Article Views and Revisions

Khoi-Nguyen Tran and Peter Christen

Research School of Computer Science
The Australian National University, Canberra, ACT 0200, Australia
{khoi-nguyen.tran,peter.christen}@anu.edu.au

Abstract. Vandalism is a major issue on Wikipedia, accounting for about 2% (350,000+) of edits in the first 5 months of 2012. The majority of vandalism are caused by humans, who can leave traces of their malicious behaviour through access and edit logs. We propose detecting vandalism using a range of classifiers in a monolingual setting, and evaluated their performance when using them across languages on two data sets: the relatively unexplored hourly count of views of each Wikipedia article, and the commonly used edit history of articles. Within the same language (English and German), these classifiers achieve up to 87% precision, 87% recall, and F1-score of 87%. Applying these classifiers across languages achieve similarly high results of up to 83% precision, recall, and F1-score. These results show characteristic vandal traits can be learned from view and edit patterns, and models built in one language can be applied to other languages.

1 Introduction

Wikipedia is the largest free and open access online encyclopedia. It is written by millions of volunteers and accessed by hundreds of millions of people each month. These kinds of large open collaborative environments are naturally attractive to vandals. A malicious modification to a Wikipedia article is available instantly to millions of potential readers. Vandalism comes in many forms, where we adopt the definitions of Priedhorsky et al. [1], repeated here for convenience: misinformation, mass delete, partial delete, offensive, spam, nonsense, and other.

Vandalism is a key issue on Wikipedia, despite the majority of vandalism being caught and repaired very quickly [1–3]. Finding and repairing these vandalisms distracts Wikipedia editors from writing articles and other important work. To lighten the burden of finding and resolving vandalism, anti-vandalism bots have been created and are operating since 2006. Although these bots use simple rules and word lists, they find the majority of obvious vandalism cases [4].

As Wikipedia grows larger and vandals adapt to anti-vandalism bots, new techniques are needed to combat vandalism. Many machine learning techniques (see Sect. 2) offer potential automated solutions. Vandalism is commonly identified from user comments in Wikipedia data dumps of the complete edit history,

J. Pei et al. (Eds.): PAKDD 2013, Part II, LNAI 7819, pp. 268–279, 2013.

where patterns in language, content, metadata, users, and others can be modelled. Various features, ranging from simple metadata to complex word analyses, are constructed for machine learning algorithms. These vandalism studies often use the English Wikipedia, but rarely the other 280+ language editions.

In this paper, we explore crosslingual vandalism detection by using a relatively unexplored data set, the hourly article view count, and the commonly used complete edit history of Wikipedia. We also combine these two data sets to observe any benefits from additional language independent features. We look at two language editions, English and German, and compare and contrast the performance of standard classifiers in identifying vandalism within a language and applied across language.

We hypothesise vandalism can be characterised by the view patterns of a vandalised articles. Vandals may be eliciting behavioural patterns before, during, and after a vandalised edit. We further hypothesise that behaviour of vandals is similar across language domains. This means models developed in one language can be applied to other languages. This can potentially reduce the cost of training classifiers for each language. We find this cross language application of vandalism models produces similarly high results as for a single language.

Our contributions are (1) novel use of the hourly article view data set for vandalism detection; (2) creation and combination of data sets with language independent features; and (3) showing the cross language applicability of vandalism models built for one language.

The rest of this paper is organised as follows. Section 2 reviews the related work. Section 3 provides statistics of the Wikipedia data sets and how to create the combined data set. Section 4 details the machine learning algorithms and their parameters. Section 5 summarises the results, providing precision, recall, F1-score, and execution times. Section 6 discusses the significance, quality, and limitations of this data set and approach. Finally, we conclude this paper in Section 7 with outlook to future work.

2 Related Work

We survey some of the most related research on vandalism detection. Vandalism is a prominent issue on Wikipedia, which arise in many research looking the dynamics of Wikipedia. One increasingly popular approach of finding vandalism is to use machine learning techniques. This approach and others are applied to a Wikipedia vandalism detection competition at the PAN workshop[1].

The complex open collaborative environment of Wikipedia has seen many studies trying to comprehend the interactions that lead to developing content. By its open nature, vandalism or more general malicious edits have occurred on every Wikipedia article [2]. Vandalism is a burden on Wikipedia, where its occurrence and work in identifying and reverting it are increasing [3]. The time spent on maintenance work, such as reverting vandalism, by Wikipedians (registered users) are increasing, which leave less time for writing articles [3].

[1] `http://pan.webis.de/`

Wikipedians have a variety of ways to deal with vandalism, which including developing and using tools to identify vandalism, such as bots [5]. Many types of vandalism can be identified clearly from visualisations of the edit history using flow diagrams [2]. Other types of vandalism require more complex analysis of the article content. Although many cases of vandalism are repaired almost immediately [1–3], the probability that an article will be vandalised is increasing over time [1].

Vandalism often has many characteristics, where use of machine learning is becoming increasingly common [6]. These machine learning techniques require building features from the Wikipedia data sets, which can range from simple metadata to more complex analysis of content, semantics, authors, and interactions. Anti-vandalism bots have been constantly monitoring Wikipedia since 2006, but the simple features and constructed rules and word lists used by the bots can be easily deceived and leave room for improvement [4].

Analysing the words used in the content of articles can provide evidence of vandalism. When comparing revisions of an article, word level features can determine whether the use of certain words will be rejected and reverted in later revisions [7]. The revision history of an article offers a distribution of words relevant to that article. This word distribution allows machines to find use of unexpected words, which is a common type of vandalism [8]. More general analyses of words and content often use natural language processing techniques, which can provide models that well surpass rule based approaches and other machine learning approaches [9]. Linguistic features from applying natural language processing can characterise vandalism and be learned by machines [10].

By combining content analyses with other information about authors and objective measures of edit quality, reputation systems can be developed to identify vandalism [11]. Without these features, spatio-temporal properties of metadata can be sufficient for machine learning algorithms to detect vandalism [12]. However, machine learning algorithms can be improved by using many features, to which some research use a range of features identified from past research studies to train algorithms [12].

In recent years, the task of identifying vandalism on Wikipedia has been turned into a competition. The PAN Workshop hosted Wikipedia vandalism detection competitions as part of its workshops in 2010 and 2011. In 2010, a vandalism corpus was created using the Amazon's Mechanical Turk to label its data set [13]. This crowdsourcing of vandalism identification proved to be successful and a larger crowdsourced corpus of over 30,000 Wikipedia edits was released in 2011, and in three languages: English, German, and Spanish [14]. This multilingual vandalism corpus uses 65 features to quantify characteristics of an edit to capture vandalism. The 2010 winner explored metadata features from edits and expanded word list features for a Random Forest classifier [15]. A post 2010 competition study combined spatio-temporal analysis of metadata [12], reputation system [11], and natural language processing features to further improve on the winning system. The 2011 winner focused on language independent features and constructed 65 features for an alternating decision tree classifier [16].

Table 1. Basic statistics of edit history data set. All revisions until start of June 2012.

Language	Content articles	Article revisions	Distinct usernames	Distinct IP addresses
English	4,000,264	305,821,091	4,020,470	25,669,884
German	1,419,217	65,732,032	447,603	5,565,475

Table 2. Basic statistics of article view data set. From January 2012 to May 2012.

Language	Articles viewed	Total views
English	2,261,593	4,567,904,954
German	805,964	1,493,732,111

3 Wikipedia Data Sets

In this section, we describe the process of generating the data sets used for vandalism classification. We use two data sets: the complete edit history of Wikipedia in English and German[2], and the hourly article view count[3]. We describe data with language codes "en" for English and "de" for German. These two raw data sets are processed as described in the subsections below.

We use the edit history data dump of 1 June 2012 for the English Wikipedia, and 3 June for the German Wikipedia. Table 1 summarises the number of articles and revisions, and distinct usernames. Content articles are strictly encyclopedic articles and do not include articles for redirects, talk, user talk, help, and other auxiliary article types. We provide count of usernames and IP addresses in Table 1 to give indication of activity in the two Wikipedias.

The raw article view data set contains all of MediaWiki projects (including Wikipedia). As of writing this paper, we have obtained all data from January to May 2012. We filter only revisions made in this time period from the edit history data. Table 2 provides some basic statistics on the raw data set filtered to view counts of English and German articles. Accordingly, we filtered the edit history data set to revisions made between January and May 2012.

3.1 Vandalised Revisions

From the raw revision data, every revision is reduced to a vector of features described in Table 3. These features are selected for their language independence and simplicity. For each revision, we analyse its comment for keywords of "vandal" and "rvv" (revert due to vandalism), indicating the occurrence of vandalism in the previous revision(s). The appropriate revisions are then marked as an occurrence of vandalism.

To align the timestamp of revisions to the corresponding article view data set, we round up the revision time to the next hour. This ensures that the hourly

[2] `http://dumps.wikimedia.org/backup-index.html`
[3] `http://dumps.wikimedia.org/other/pagecounts-raw/`

Table 3. Description of edit history data set

Attribute	Description
Article title	Unique identifier of a Wikipedia article.
Hour timestamp	The timestamp of this revision. In the format of YYYYMMDD-HH0000. The minutes and seconds are used to round up to the next hour.
Anonymous edit	The editor of this revision is considered to be anonymous if an IP address is given. 0 for an edit by a registered user, and 1 for an edit by an anonymous user.
Minor revision	Revisions can be flagged as minor edits. 0 for normal revision, and 1 for minor revision.
Size of comment (bytes)	The size of the given comment of this revision.
Size of article text (bytes)	The size of the complete article of this revision.
Vandalism	This revision is marked as vandalism by analysing the comment of the following revision(s). 0 for not vandalism, and 1 for vandalism.

Table 4. Description of article view data set

Attribute	Description
Project name	The name of the MediaWiki project, where we are interested in Wikipedia projects in English ("en") and German ("de").
Hour timestamp	In the format of YYYYMMDD-HH0000, where YYYY for year; MM for month; DD for day of the month; HH for 24-hour time (from 00 to 23); and minutes and seconds are not given.
Article title	The title of the Wikipedia article. Article may not exist as the data set is derived from Web server request logs.
Number of requests	The number of requests in that hour. Not unique visits by users.
Bytes transferred	The total number of bytes transferred from the requests.

article views references the correct revision when combining the two data sets. The alignment is performed on all revisions and should not affect classification.

We emphasise that user labelling of Wikipedia vandalism is noisy and incomplete. Some research provides solutions to this problem such as active learning [8], but a fully automated approach have inherent limitations as human involvement is necessary for some cases of vandalism [17]. We find about 2% of revisions between January to May 2012 contain vandalism. This is consistent with studies looking at these keywords [3], but less than the 4-7% reported in other studies looking at vandalism beyond user labelling [1, 11, 13].

3.2 Article Views

The raw article view data set is structured by views of article aggregated by hour. We perform a simple transformation and filtering of articles seen in the revisions data set above. The resulting features are summarised in Table 4.

We also extract the redirect articles from the revisions data set and change all access to redirect articles to the canonical article. These extra view counts are aggregated accordingly.

These article views are important to seeing the impact of vandalism on Wikipedia [1]. With the average survival time of vandalism being 2.1 days [3], this leaves many hours for unsuspecting readers to encounter vandalised content. However, the behaviour of vandals may also be seen in a change in access patterns, which may be from vandals checking on their work, or that article drawing attention from readers and their peers.

A previous research study [1] (before the release of this data set) derived article views from the full Wikipedia server logs. This provides a much finer time unit for analysis, but with a huge increase in data to process. With the time unit of hours, this data set may provide coarse patterns of behaviours, but with manageable data size.

There are few research studies that use this data set. Most research has developed tools for better access to this huge data resource and to provide simple graphs for topic comparison. One relevant study [18] use this data set to compare access to medical information on seasonal diseases like the flu. Access patterns in this data set reflect the oncoming of seasonal diseases. Wikipedia is accessed more than other online health information providers, and is a prominent source of online health information. Although vandalism is not covered, the seasonal access patterns elude to potential targets of vandalism.

To determine whether these article views occurred when articles are in a vandalised state, we scan the edit history data set and label all article views of observed vandalised or non-vandalised revisions. The unknown views from revisions made before January 2012, or articles without revisions in this 5 month period under study, are discarded. Thus, we have an article view data set labelled with whether the views are of vandalised revisions. The resulting size of the data is identical to the combined data set in the following subsection. This labelled article view data set allows us to determine whether view patterns can be used to predict vandalism.

From this resulting combined set, we split the "Hour timestamp" attribute into an "hour" attribute. This allows the machine learning algorithm to learn daily access patterns. In future work, we intend to experiment with monthly and yearly access patterns when we have obtained enough raw data.

3.3 Combined Data Set

The combined data set is the result of merging of two time series data sets for each language. The data set is constructed by adding features from the labelled revisions data set to the labelled article view data set by repeating features of the revisions. Thus for every article view, we have information on whether a vandalised revision was viewed and what the properties of that revision are.

We use the "hour" attribute split from the timestamp in the article views data set. Thus, we have the following 8 features in our combined data set: **hour, size**

Table 5. Statistics of the various data sets. With percentage of vandalism.

Data set	Vandalised revisions	Article views	Combined (train)	Combined (test)
English	17,159,583	525,382,429	271,584,092	99,611,391
(vandal)	(2.08%) 356,618	-	(2.34%) 6,367,602	(2.04%) 2,033,838
German	3,731,714	284,932,083	139,967,644	55,010,679
(vandal)	(0.10%) 3,889	-	(0.06%) 86,534	(0.07%) 40,143

of comment, size of article, anonymous edit, minor revision, number of requests, bytes transferred, and **vandalism** (class label).

These features are language independent and capture the metadata of revisions commonly used, and access patterns. Note that we remove the article name as they are not necessary in evaluating the quality of classification. For example, access patterns of vandalised articles may be similar to other vandalised articles, regardless of the name of articles. For future work, we may identify the articles classified and further analyse to determine genuine cases of vandalism unlabelled or overlooked by editors.

To apply the classification algorithms, we split the combined data set by date into a training set (January to April) and a test set (May). The statistics of the data sets in this section are shown in Table 5 for comparison.

4 Cross Language Vandalism Prediction

We use the Scikit-learn toolkit [19], which provides many well-known machine learning algorithms for science and engineering. We selected the following supervised machine learning algorithms from the toolkit:

- Decision Tree (DT)
- Random Forest (RF)
- Gradient Tree Boosting (GTB): binomial deviance as the loss function.
- Stochastic Gradient Descent (SGD): logistic regression as the loss function.
- Nearest Neighbour (NN): KDTree data structure.

We experimented with different settings available for the classifiers above, but we found there is little to no variance in the results. This is likely because all classifiers converged with the already large number of observations given.

From Table 5, we see the data set is highly unbalanced, which is unsuitable for some of our classifiers. We resolved this problem by undersampling the non-vandalism observations to match the number of vandalism observations. We apply this to all three data sets. Thus, we built a balanced subset of the training and testing data.

We repeated the application of the classifiers to the balanced data to observe any effects from the random samples of non-vandalism observations. We found all classifiers seem to have converged with the already large number of observations in the balanced subset.

Table 6. Classification results of the revisions data set

Model-Language	Precision					Recall					F1-Score				
	DT	RF	GTB	SGD	NN	DT	RF	GTB	SGD	NN	DT	RF	GTB	SGD	NN
en-en	0.78	**0.84**	**0.84**	**0.84**	0.69	0.78	0.83	**0.84**	**0.84**	0.69	0.78	0.83	**0.84**	**0.84**	0.69
de-de	0.75	**0.84**	**0.84**	0.69	0.69	0.74	0.83	**0.84**	0.51	0.68	0.74	0.83	**0.84**	0.36	0.68
de-en	0.70	0.81	**0.82**	0.64	0.59	0.70	0.80	**0.82**	0.51	0.57	0.70	0.80	**0.81**	0.35	0.56
en-de	0.76	0.82	**0.83**	**0.83**	0.58	0.76	0.82	**0.83**	**0.83**	0.56	0.76	0.82	**0.83**	**0.83**	0.54

Table 7. Classification results of the article views data set

Model-Language	Precision					Recall					F1-Score				
	DT	RF	GTB	SGD	NN	DT	RF	GTB	SGD	NN	DT	RF	GTB	SGD	NN
en-en	**0.82**	0.55	0.78	0.62	0.69	**0.80**	0.53	0.73	0.50	0.69	**0.80**	0.48	0.72	0.35	0.69
de-de	**0.81**	0.69	0.70	0.25	0.69	**0.74**	0.69	0.70	0.50	0.68	**0.72**	0.69	0.70	0.33	0.68
de-en	0.55	0.63	**0.68**	0.25	0.59	0.50	0.63	**0.68**	0.50	0.57	0.35	0.62	**0.68**	0.33	0.56
en-de	0.60	0.51	**0.62**	0.54	0.58	0.55	0.50	**0.62**	0.50	0.56	0.48	0.42	**0.62**	0.34	0.54

Table 8. Classification results of the combined data set

Model-Language	Precision					Recall					F1-Score				
	DT	RF	GTB	SGD	NN	DT	RF	GTB	SGD	NN	DT	RF	GTB	SGD	NN
en-en	**0.86**	**0.87**	**0.85**	0.84	0.69	**0.84**	**0.87**	**0.85**	0.84	0.69	**0.83**	**0.87**	**0.85**	0.84	0.69
de-de	0.81	0.84	**0.88**	**0.72**	0.69	0.74	0.82	**0.87**	0.51	0.68	0.72	0.82	**0.87**	0.35	0.68
de-en	0.65	0.73	**0.83**	0.60	0.59	0.53	0.68	0.82	0.50	0.57	0.42	0.66	**0.82**	0.34	0.56
en-de	0.70	0.77	0.82	0.83	0.58	0.58	0.75	0.82	0.83	0.56	0.51	0.75	0.82	0.83	0.54

We also tried to train a Support Vector Machine (SVM) classifier, but we are unable to obtain results because of the different order in magnitude of training time. We experimented with very few number of samples (0.1-1% of the data set) to obtain results for SVM within a reasonable time frame. However, we found all classifiers above and including the SVM performed poorly with the small number of observations.

For cross language vandalism prediction, we first train classification models for our two languages: English and German. These models are then evaluated on the testing set for the same language, then to the testing set of the other language. This may seem odd with the independent nature of language domains. However, our data sets capture language independent features of Wikipedia. This cross language application of models allows a generalisation of editing and viewing behaviour across Wikipedia.

This cross language application of models has seen successful applications in the research area of cross language text categorisation [20, 21]. When considering text, cultural knowledge of the target language is needed to inform classifiers. The advantage of cross language application of models is that one model can be used for multiple languages, saving resources developing models for each language. This is particularly relevant to Wikipedia with its large range of languages. This research allows the potential generalisation of the concentration of vandalism research in English to other languages without additional inputs.

Table 9. Approximate execution time of classifiers in seconds

Time Taken (s)	DT	RF	GTB	SGD	NN
Training (en)	750	550	800	5	20
Training (de)	3	4	15	1	1
Testing (en-en)	5	16	3	0.5	150
Testing (de-de)	0.5	0.5	0.5	0.5	2
Testing (de-en)	2	7	5	2	90
Testing (en-de)	0.5	0.5	0.5	0.5	4

5 Experimental Results

The classification results are presented in Tables 6, 7, and 8. These are the total obtained scores from classification of the two classes: vandalism and non-vandalism. They present the classification results of a classifier trained in one language and applied to another. For example, "en-de" means the classification model is trained on the English training set, then applied to the German testing set. The highest classification scores of the classifier group are highlighted in bold font in Tables 6 and 7. For the combined data set, the highest scores and scores that outperformed the individual data sets are highlighted in bold font in Table 8. The approximate execution times, gathered and rounded from multiple runs, are summarised in Table 9.

For the monolingual application of classification models in the single data sets, the tree based methods generally have better performance. In particular GTB and RF for the revisions data set, and DT for the views data set. They are also the most expensive models to train.

The crosslingual application showed similar, but generally weaker, performance across all measures. GTB and RF continue to show generally better performance than the other classifiers. Interestingly, SGD performed best in the monolingual and crosslingual cases when trained on the English revisions data, suggesting English may offer more patterns to detect vandalism. This is encouraging because SGD is the fastest algorithm to train. The crosslingual application of models is not detrimental in most cases for all data sets, but with similar performance to the monolingual case. This suggests cross language classification of vandalism is feasible with a variety of data sets.

In the combined data set, we see improvements to the classification scores, but mainly in the monolingual case. GTB continues to show high performance with improvements from the additional features. In general the combination of the data sets does not provide a significant advantage to the classifiers. The classifiers seem to do as well on the combined data set compared to individual data sets, but not much better. This suggests the classifiers are learning the best models from each data set, but improvements are not common.

The monolingual classification scores of the revisions data set in Table 6 are comparable and better than many state-of-the-art systems. Note that the data sets used in various research studies are often constructed differently, and so

care is needed when comparing different studies. From overviews of the PAN Wikipedia Vandalism Detection competition [14, 22], our results show better performance than many of entries, while using fewer features. The competition showcased multilingual entries in 2011, but no cross language application of models is seen. White and Maessen [23] presents an entry into the 2010 PAN vandalism competition and collated results from other Wikipedia vandalism research. We find our results for monolingual classification to generally have higher precision, recall, and F1-score.

6 Discussion

Vandalism is an important cross language issue on Wikipedia as more people contribute to and use Wikipedia as a resource in many different languages. The current research on vandalism shows promising technologies to automatically detect and repair vandalism. However, these research studies largely concentrate on the English Wikipedia. The generalisation of these studies to other languages may not always be possible because of the independence of language domains, and the peculiarities in languages. Multilingual vandalism research is appearing, aided by construction of multilingual vandalism data sets, such as those by the PAN workshop. The cross language vandalism detectors are ideal as models develop in one language can be applied to other languages.

The advantages of the presented data sets are the simple to extract language independent features. These few features with the application of baseline classification algorithms outperform many past research studies. The combination of editing and viewing patterns shows some increase in performance, but generally allows classifiers to adapt to the best predictive features from both data sets individually. The article view data set may be too coarse to predict vandalism at the hourly level, but we found some classifiers can find patterns of vandalism as well, or better than the revisions data set in some cases.

Some limitations of our approach include using few features, not analysing the content, and the necessity of the revisions data set to label the article views data set. The rich number of features used in other studies allows classifiers to learn more patterns of vandalism. This can often improve performance, but we find these data sets can be difficult to generate, especially when deploying solutions in bots. We have ignored the content of revisions, where word analysis may show the clear cases of unlabelled cases of vandalism. However, this is simply not feasible on a large scale required for Wikipedia and its many languages.

Our data set offers indications of vandalism that can be investigated with more complex techniques. The article views data set alone is not sufficient for vandalism detection and requires labelling from the revisions data set. However, by building labelled article views data sets, unlabelled articles can be incorporated and learned in a semi-supervised setting. Despite these limitations, we have shown cross language application of vandalism models is feasible, and view patterns can be used to predict vandalism and may offer improvements to classifiers.

7 Conclusion

We have presented data sets for vandalism detection and demonstrated the application of various machine learning algorithms to detect vandalism within one language and across languages. We developed three data sets from the hourly article view count data set, complete edit history of Wikipedia, and their combination. We looked at two language editions of Wikipedia: English and German. Within the same language, these baseline classifiers achieve up to 87% precision, 87% recall, and an F1-score of 87%. The cross language application of these classifiers achieved similarly high results of up to 83% precision, recall, and F1-score. We find Gradient Tree Boosting showed generally best performance in predicting vandalism, despite being the most time consuming algorithm. These results show the view and edit behaviour of vandals is similar across different languages. The implication of this result is that vandalism models can be trained in one language and applied to other languages.

In future work, we could extend the time span of the data set and apply to other languages. This would provide further evidence for the general applicability of classification models cross language to detect vandalism using this combined data set. We may add further features to enrich the data set and explore other balancing techniques. We could improve the baseline classifiers by building classifiers more suited to this data set. In the long term, we plan to have this system able to generate the data set in near real time and predict possible cases of vandalism for closer analysis.

References

1. Priedhorsky, R., Chen, J., Lam, S.T.K., Panciera, K., Terveen, L., Riedl, J.: Creating, destroying, and restoring value in wikipedia. In: Proceedings of the 2007 International ACM Conference on Supporting Group Work, GROUP 2007, pp. 259–268. ACM, New York (2007)
2. Viégas, F.B., Wattenberg, M., Dave, K.: Studying cooperation and conflict between authors with history flow visualizations. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2004, pp. 575–582. ACM, New York (2004)
3. Kittur, A., Suh, B., Pendleton, B.A., Chi, E.H.: He says, she says: conflict and coordination in wikipedia. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2007, pp. 453–462. ACM, New York (2007)
4. Smets, K., Goethals, B., Verdonk, B.: Automatic vandalism detection in wikipedia: Towards a machine learning approach. In: AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy, pp. 43–48 (2008)
5. Panciera, K., Halfaker, A., Terveen, L.: Wikipedians are born, not made: a study of power editors on wikipedia. In: Proceedings of the ACM 2009 International Conference on Supporting Group Work, GROUP 2009, pp. 51–60. ACM, New York (2009)
6. Potthast, M., Stein, B., Gerling, R.: Automatic vandalism detection in wikipedia. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 663–668. Springer, Heidelberg (2008)

7. Rzeszotarski, J., Kittur, A.: Learning from history: predicting reverted work at the word level in wikipedia. In: Proc. of the ACM 2012 Conf. on Computer Supported Cooperative Work, CSCW 2012, pp. 437–440. ACM, New York (2012)
8. Chin, S.C., Street, W.N., Srinivasan, P., Eichmann, D.: Detecting wikipedia vandalism with active learning and statistical language models. In: Proc. of the 4th Workshop on Information Credibility, WICOW 2010, pp. 3–10. ACM (2010)
9. Wang, W.Y., McKeown, K.: "got you!": Automatic vandalism detection in wikipedia with web-based shallow syntactic-semantic modeling. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Beijing, China. Coling 2010 Organizing Committee, pp. 1146–1154 (August 2010)
10. Harpalani, M., Hart, M., Singh, S., Johnson, R., Choi, Y.: 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, vol. 2, pp. 83–88 (2011)
11. Adler, B., de Alfaro, L., Pye, I.: Detecting wikipedia vandalism using wikitrust. Notebook Papers of CLEF 1, 22–23 (2010)
12. West, A.G., Kannan, S., Lee, I.: Detecting wikipedia vandalism via spatio-temporal analysis of revision metadata? In: Proceedings of the Third European Workshop on System Security, EUROSEC 2010, pp. 22–28. ACM, New York (2010)
13. Potthast, M.: Crowdsourcing a wikipedia vandalism corpus. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, pp. 789–790. ACM, New York (2010)
14. Potthast, M., Holfeld, T.: Overview of the 2nd international competition on wikipedia vandalism detection. In: Notebook for PAN at CLEF (2011)
15. Velasco, S.: Wikipedia vandalism detection through machine learning: Feature review and new proposals. In: Lab Report for PAN-CLEF 2010 (2010)
16. West, A.G., Lee, I.: Multilingual vandalism detection using language-independent & ex post facto evidence - notebook for pan at clef 2011. In: Petras, V., Forner, P., Clough, P.D. (eds.) CLEF (Notebook Papers/Labs/Workshop) (2011)
17. Wu, Q., Irani, D., Pu, C., Ramaswamy, L.: Elusive vandalism detection in wikipedia: a text stability-based approach. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM 2010, pp. 1797–1800. ACM, New York (2010)
18. Laurent, M., Vickers, T.: Seeking health information online: does wikipedia matter? Journal of the American Medical Informatics Association 16(4), 471–479 (2009)
19. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
20. Rigutini, L., Maggini, M., Liu, B.: An em based training algorithm for cross-language text categorization. In: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 529–535 (September 2005)
21. Liu, Y., Dai, L., Zhou, W., Huang, H.: Active learning for cross language text categorization. In: Tan, P.-N., Chawla, S., Ho, C.K., Bailey, J. (eds.) PAKDD 2012, Part I. LNCS, vol. 7301, pp. 195–206. Springer, Heidelberg (2012)
22. Potthast, M., Stein, B., Holfeld, T.: Overview of the 1st international competition on wikipedia vandalism detection. In: Braschler, M., Harman, D., Pianta, E. (eds.) CLEF (Notebook Papers/LABs/Workshops) (2010)
23. White, J., Maessen, R.: Zot! to wikipedia vandalism - lab report for pan at clef 2010. In: CLEF (Notebook Papers/LABs/Workshops) (2010)