# Wikipedia Vandalism Detection

Santiago M. Mola-Velasco
Supervised by Paolo Rosso
NLE Lab. - ELiRF, Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
Camino de Vera s/n. 46022 Valencia, Spain
{smola,prosso}@dsic.upv.es

## ABSTRACT

Wikipedia is an online encyclopedia that anyone can access and edit. It has become one of the most important sources of knowledge online and many third party projects rely on it for a wide-range of purposes. The open model of Wikipedia allows pranksters, lobbyists and spammers to attack the integrity of the encyclopedia and this endangers it as a public resource. This is known in the community as vandalism.

A plethora of methods have been developed within the Wikipedia and the scientific community to tackle this problem. We have participated in this effort and developed one of the leading approaches. Our research aims to create a fully-working antivandalism system and get it working in the real world.

## Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering

## General Terms

Algorithms, Experimentation, Measurement, Performance

## Keywords

Wikipedia vandalism detection, machine learning, natural language processing, reputation

## 1. INTRODUCTION

During the last 10 years, we have witnessed the birth and rise of the biggest encyclopedia ever created: Wikipedia. The unique aspect of this online encyclopedia is that anyone can freely access and edit it. Thanks to this collaboration model, there are actively maintained editions in 240 languages, a English edition with more than 3 million articles and over 13 million registered users. Its average quality has been demonstrated to be as good as well-established encyclopedias [9] and it is used as a encyclopedic knowledge source, not only by people who use it directly, but also by third party projects such as knowledge databases, e.g. DBPedia[1], definitions for dictionaries, e.g. Google[2], and educational projects for developing countries, e.g. Wikipedia 1.0[3].

While being an open community where anyone can participate and contribute is the essence of Wikipedia and it is at the core of its success, it also generates problems that endanger the proper development of the project. One of these problems is that pranksters, lobbyists and spammers target Wikipedia for their dubious purposes. This has a negative impact on the encyclopedia itself and, indirectly, on every third party project that uses it. In this research, we focus in *vandalism*, which is defined by Wikipedia as follows [23]: *Vandalism is any addition, removal, or change of content made in a deliberate attempt to compromise the integrity of Wikipedia.*

Vandalism is a highly subjective concept. At this point we do not concern ourselves with the delimitation of the concept and work with corpora annotated by humans, who judge on a case-by-case basis, as our ground truth. There are many kinds of vandalism [21, 18, 7, 23], Wikipedia contributors identify 20 categories [23], of which we consider the following[4]:

**Blanking** Removing all or significant parts of a page's content.

**Edit summary vandalism** Making offensive edit summaries in an attempt to leave a mark that cannot be easily expunged from the record.

**Hidden vandalism** Any form of vandalism not visible in the final article but visible during editing.

**Image vandalism** Uploading shock images or inappropriately placing explicit images.

**Link vandalism** Adding or changing internal or external links on a page to disruptive, irrelevant, or inappropriate targets while disguising them with mislabeling.

**Illegitimate page creation** Creating new pages with the sole intent of malicious behaviour.

[1]See `http://dbpedia.org/`.

[2]Google provides the `define` command to provide definitions through its search engine. See `http://www.google.com/help/features.html`.

[3]See `http://meta.wikimedia.org/wiki/Wikipedia_1.0`.

[4]We excluded *malicious account creation*, *abuse of tags*, *avoidant vandalism*, *repeated upload of copyrighted material* and *gaming the system* because they are beyond the scope of our research.


*WWW 2011*, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

**Page lengthening** Adding very large amounts of content to a page so as to make the page's load time abnormally long.

**Page-move vandalism** Changing the names of pages to disruptive, irrelevant o inappropriate names.

**Silly vandalism** Adding profanity, graffiti or nonsense to pages.

**Sneaky vandalism** Vandalism that is harder to spot, or that otherwise circumvents detection.

**Spam external linking** Adding links to irrelevant sites after having been warned.

**Template vandalism** Modifying the wiki language or text of a template in a harmful manner.

The frequency of each vandalism type has not been explored thoroughly. However, based on a previous study with a small sample [7] and personal experience, we can claim that *silly* and *sneaky* vandalism constitutes the large majority of vandalism. Being the former easiest to detect along with *blanking* and the later the most hard to detect and serious cases.

Currently, Wikipedia relies on a number of dedicated volunteers that check every change, detect vandalism and revert it. This task is very time-consuming and the massive volume of changes makes human intervention insufficient. We can get a rough idea of this volume with the fact that only in the English edition of Wikipedia, there were 10 million edits between August 20 and October 10 2010, which makes almost 200 thousand edits per day on average[5]. It is estimated that the vandalism rate is around 7% [15], so we can assume that there are about 14 thousand vandalism edits every day. The Wikipedia community develops bots, i.e. autonomous systems, that check every edit in real time, try to spot vandalism, and revert it [5, 19]. In recent years, the scientific community has become involved in the research and development of techniques for automated vandalism detection.

We started working on this problem in early 2010 participating in the 1st International Competition on Wikipedia Vandalism Detection, organized by PAN 2010 Evaluation Lab.[6], held in conjunction with CLEF 2010 [17]. After exploring a simple approach and getting the first place in the competition, we have conducted research with the purpose of pushing Wikipedia vandalism detection state-of-the-art to higher levels, aiming to build an automated system that can work autonomously.

The rest of the article is structured as follows: in Section 2 we give a general description of the Wikipedia vandalism detection task. Section 3 overviews the state-of-the-art. In Section 4 we detail our approach, previous and ongoing work. Section 5 summarizes our results and in Section 6 we draw some conclusions.

## 2. WIKIPEDIA VANDALISM DETECTION

To define the vandalism detection task, we have to define some key concepts of MediaWiki, i.e. the wiki engine used by Wikipedia. An article is composed of a sequence of revisions, commonly referred to as the article history. A revision is the state of an article at a given time in its history and is composed of the textual content and metadata describing the transition from the previous revision. Revision metadata contains, among others, the user who performed the edit, a comment explaining the changes, a timestamp, etc. An edit is a tuple of two consecutive revisions and should be interpreted as the transition from a given revision to the next one. Wikipedia vandalism detection is a one-class classification task. The goal is, given any edit, determine whether it is destructive or not.

Evaluating a vandalism detection system requires a corpus of pre-classified edits. Four different corpora have been reported in the literature [16, 22, 7, 15]. We have chosen the PAN-WVC-10 corpus[7] [15]. This corpus contains 32,452 edits on 28,468 articles, where 2,391 are labelled as vandalism. These edits were collected during a week and the distribution of the vandalism class is claimed to correspond to the actual distribution in Wikipedia. This is the only corpus where distribution has been considered and where each edit is annotated by more than one human. The PAN-WVC-10 was created for the 1st International Competition on Wikipedia Vandalism Detection as part of the PAN Evaluation Lab, held in conjunction with CLEF 2010[8]. It is the most recent and widely used corpus.

Performance is measured using standard Information Retrieval measures. In this case, Precision, Recall and Receiver Operating Characteristic (ROC). In this paper, we refer to Area Under Precision-Recall Curve as the reference measure.

## 3. RELATED WORK

The main approach of systems being used in Wikipedia for automatic vandalism detection is the use of heuristics aimed at detecting very specific kinds of vandalism. This heuristics include the amount of text inserted or deleted, the amount of uppercase letters and the frequency of vulgarisms detected via regular expressions [5, 19]. ClueBot is one of the most prominent systems currently in use. In one study it is found to have 100% precision but very low recall, 49% for deletions and 4% for insertions [16], another study concluded with similar estimations [20]. Besides its low recall, these systems are difficult to maintain because of the need of creating lists of regular expressions and manually adjusting weights and thresholds. Another problem is their limited capacity for being applied to different languages.

Following the ideas of antivandalism bots, Potthast et al. [16] cast the problem as a machine learning one-class classification problem and manually inspect 301 cases of vandalism to create a set of features based in the text and metadata of the edit. These features include the uppercase ratio, frequency of vulgarisms and personal pronouns, size change in the article, whether the editor is anonymous or not, among others. Most research on vandalism detection have explored features based on text [16, 20, 8, 11, 7].

Many works have considered metadata in their features [5, 19, 16, 8], but their exploitation of this information was limited to few aspects of it such as if the editor is anonymous

[5]See more statistics at `http://en.wikipedia.org/wiki/User:Katalaveno/TBE` and `http://en.wikipedia.org/wiki/Wikipedia:Statistics`

[6]See `http://pan.webis.de/`.

[7]Available at `http://www.uni-weimar.de/cms/medien/webis/research/corpora/pan-wvc-10.html`.

[8]See `http://pan.webis.de/`

or not. West et al. [22] demonstrated that metadata alone is much more discriminative than previously thought.

A completely different approach are reputation systems, pioneered in in [24, 13, 1]. West et al. [22] already applied the idea of reputation to editors and articles, as well as countries where the editors are. Adler et al. [2] demonstrate that a mixture of user and text reputation and simple metadata features results in good performance, opening a third way beyond content and metadata.

A promising approach is the use of the frequency of vandalism in a given article as a feature [6] and other features that can characterize the *a priori* probability of an article to be vandalized given its content. None of the state-of-the-art do not cover this perspective in depth yet.

The first systematic review and organization of features appears by Potthast et al. [17] as part of the PAN 2010 Evaluation Lab. The authors conclude their analysis by building a meta-classifier using the predictions of the nine participants in the competition. This meta-classifier performed significantly better than the best single participants, which suggests that the success in vandalism detection relies on the combination of a wide variety of features from all approach: content, metadata and reputation. Precision-Recall curves for PAN 2010 participants are shown in Figure 1.

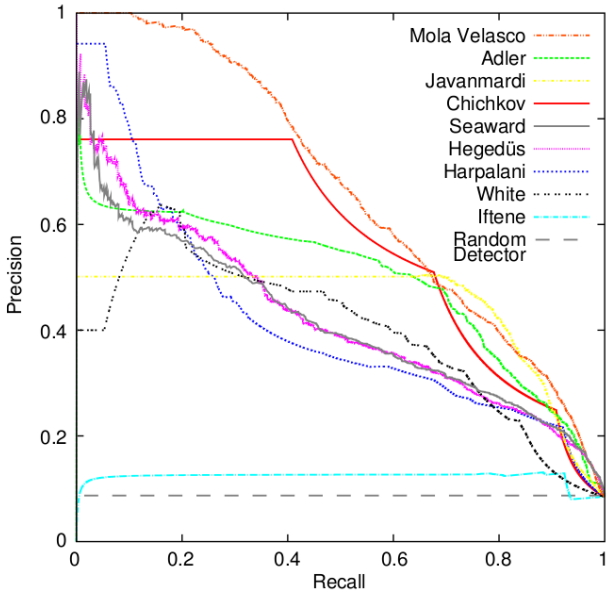


**Figure 1: Precision-Recall curves for every system that participated in the PAN 2010 Evaluation Lab. Figure extracted from the competition overview [17].**

The work by Mola-Velasco [14], ranking in the first place of PAN 2010 Evaluation Lab. extends the content and metadata-level features proposed by Potthast et al. [16]. These 21 features comprehensively model the content of the edit, including features involving the use of language, formatting of text, compressibility, spelling, and the size of the edit.

## 4. OUR APPROACH

Our first approach [14] extended the Potthast et al. [16] approach and proposed a set of language-independent features and a set of language-dependent ones. Language-independent features are described in Table 1. Language-dependent features consisted of measuring frequency of certain categories of words and impact, i.e. the percentage by which the edit increases the frequency of such words in the article. The categories of words considered are described in Table 2. As classifier we used Random Forest [4] with iterations fixed to 500. This approach has been proven to achieve competitive performance, with an Area Under Precision-Recall Curve (AUC-PR) of 0.7332, and it is the foundation of all our subsequent work.

**Table 1: Language-independent features from Mola-Velasco.**

| Feature | Description |
|---|---|
| Anonymous | Whether the editor is anonymous or not. |
| Comment length | Length in characters of the edit summary. |
| Upper to lower ratio | Uppercase to lowercase letters ratio. |
| Upper to all ratio | Uppercase letters to all letters to ratio. |
| Digit ratio | Digit to all characters ratio. |
| Non-alphanumeric ratio | Non-alphanumeric to all characters ratio. |
| Character diversity | Measure of different characters compared to the length of inserted text, given by the expression $length^{\frac{1}{different\ chars}}$. |
| Character distribution | Kullback-Leibler divergence of the character distribution with respect the expectation. |
| Compressibility | Compression rate of inserted text using the LZW algorithm. |
| Size increment | Absolute increment of size. |
| Size ratio | Size of the new revision relative to the old revision. |
| Average term frequency | Average relative frequency of inserted words in the new revision. |
| Longest word | Length of the longest word in inserted text. |
| Longest character sequence | Longest consecutive sequence of the same character. |

In our previous work, we found that there are cases of vandalism that are very hard to spot without knowledge of the encyclopedic language and the topic covered in the article. In order to solve this, we propose a topic-sensitive and language-independent method. With this method, given an edit, we retrieve a set of related articles. We use this set of articles to build a language model and measure the variation that the edit produces in the Kullback-Leibler divergence [12] between the article and the set of related articles. We used the links included in the articles to get the set of related ones and then, tried unigram, bigram and skip-gram [10] language models. This improved the performance of our classifier from an AUC-PR of 0.7332 up to 0.7533, a promis-

**Table 2: Sets of words for language-dependent features in Mola-Velasco.**

| Type | Description |
|---|---|
| Vulgarisms | Vulgar and offensive words. |
| Pronouns | First and second person pronouns, including slang spellings. |
| Biased | Colloquial words with high bias. |
| Sex | Non-vulgar sex-related words. |
| Bad | Hodgepodge category for colloquial contractions and common typos. |
| All | A meta-category, containing vulgarisms, pronouns, biased, sex-related and bad words. |
| Good | Words rarely used by vandals, mainly wiki-syntax elements. |

ing result for a completely language-independent method that accounts for the use of language in each article.

Our published work, so far, includes content-level features and others based on metadata. We considered that the next type of feature to explore should be reputation and a wider range of metadata features. In our most recent work [3], we initiated a joint-effort with the authors of two of the leading approaches on vandalism detection: Adler et al. [2] and West et al. [22]. We combined metadata, content-level language-dependent, content-level language-independent and reputation features. Performance for this combination as well as for each category of features was measured, obtaining a significantly better classifier than any of the separate approaches, with an AUC-PR of 0.8183. Note that our topic-sensitive method is not included in this work yet.

During his PhD, Mola-Velasco will work on improving the PAN-WVC-10 corpus. Improvements will include a) using insights from the predictions of state-of-the-art classifiers and do manual revision of false positives and negatives to detect errors, b) compute reputation values for annotators in order to improve the gold annotations and provide confidence values on them and c) create an online crowdsourcing platform to provide further and more fine-grained labels. Using this platform, combined with Amazon Mechanical Turk, a Spanish corpus equivalent to PAN-WVC-10 will be built. This will be an important step to prove, beyond theory, the language-independence of the proposed methods.

There is ongoing work to create a new method to measure the persistence of each token in Wikipedia. This is done by counting how many times each token has been inserted and deleted in the whole Wikipedia history. These counts can be used to evaluate the persistence of each token and give a measure of the average token-persistence in an article. Our hypothesis holds that this method could make obsolete the use of manually compiled lists of vulgarisms and other words, replacing them with a fully language-independent method.

Besides exploring new features for vandalism detection, the next stage of this research will be to apply it to information quality assestment. This is, not only providing a classifier to discriminate between vandalism and regular edits, but also give a quality score to every revision. This should be the natural evolution of vandalism detection and would be specially useful for third party services who need to pick the best revision of each article from the latest available.

Finally, work have already started integrating our methods in an already existing antivandalism system. This will allow the Wikipedia community to get the benefits of our work with a seamless transition, using the same tools they do right now. This will allow us to get feedback from them, in terms of their perceived performance and usefulness.

# 5. RESULTS

In Table 3 we present a summary of our previous work. The *Mola-Velasco* result denotes our first approach. *Adler et al.* and *West et al.* are provided for for comparison. *Combined* denotes the performance of these three classifiers combined. Finally, *Mola-Velasco + topic* denotes our first approach, adding our topic-sensitive language-independent method.

**Table 3: Area Under Precision-Recall Curve (AUC-PR) of Random Forest classifiers built with our base feature set, Adler et al., West et al., the combination of these three, and our base feature combined with our proposed topic-sensitive method.**

| Classifier | AUC-PR |
|---|---|
| Adler et al. | 0.6105 |
| *Mola-Velasco* | 0.7332 |
| West et al. | 0.5253 |
| *Combined* | 0.8183 |
| *Mola-Velasco + topic* | 0.7541 |

# 6. CONCLUSIONS

Wikipedia vandalism is a serious threat to the integrity of this encyclopedia. Our current work explores the limits of the state-of-the-art and improves them, considering a wide range of features for vandalism discrimination, emphasizing on the importance of language-independence of our methods. Our ongoing work is focused on new language-independent methods, the improvement of existing corpus and the creation of a Spanish corpus. This should catalyze in the mid-term into a working system supporting a wide range of languages.

Our intention is to release all our work as open source in order to comply with the open and collaborative philosophy of Wikipedia, as well as getting our approach working in the real world.

# 7. ACKNOWLEDGMENTS

Thanks to Sandra García Blasco and Alberto Barrón Cedeño for their feedback and support. B. Thomas Adler, Luca de Alfaro and Andrew G. West for their collaboration. Sara Javanmardi for her insights on our previous work. Martin Potthast for his support during the PAN 2010 competition. This work is supported by the MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i).

# 8. REFERENCES

[1] B. Adler and L. de Alfaro. A Content-Driven Reputation System for the Wikipedia. In *WWW 2007: Proceedings of the 16th International World Wide Web Conference*. ACM Press, 2007.

[2] B. Adler, L. de Alfaro, and I. Pye. Detecting Wikipedia Vandalism using WikiTrust. In M. Braschler and D. Harman, editors, *Notebook Papers of CLEF 2010 LABs and Workshops, 22-23 September, Padua, Italy*, Sept. 2010.

[3] B. T. Adler, L. de Alfaro, S. M. Mola-Velasco, P. Rosso, and A. G. West. Wikipedia Vandalism Detection: Combining Natural Language, Metadata, and Reputation Features. In A. Gelbukh, editor, *CICLing 2011*, volume 6609 of *LNCS*, Tokyo, Japan, February 2011. Springer.

[4] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[5] J. Carter. ClueBot and Vandalism on Wikipedia. 2010. `http://www.acm.uiuc.edu/~carter11/ClueBot.pdf`.

[6] D. Chichkov. Submission to the 1st International Competition on Wikipedia Vandalism Detection. In M. Braschler and D. Harman, editors, *Notebook Papers of CLEF 2010 LABs and Workshops, 22-23 September, Padua, Italy*, USA, Sept. 2010.

[7] S.-C. Chin, W. N. Street, P. Srinivasan, and D. Eichmann. Detecting Wikipedia Vandalism with Active Learning and Statistical Language Models. In *WICOW '10: Proceedings of the Fourth Workshop on Information Credibility on the Web*, Apr 2010.

[8] G. Druck, G. Miklau, and A. McCallum. Learning to Predict the Quality of Contributions to Wikipedia. In *WikiAI'08: Proceedings of the Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy*, pages 7–12. AAAI Press, 2008.

[9] J. Giles. Internet encyclopaedias go head to head. *Nature*, 438:900–901, Dec. 2005.

[10] D. Guthrie, B. Allison, W. Liu, L. Guthrie, and Y. Wilks. A Closer Look at Skip-gram Modelling. In *Proceedings of the Fifth international Conference on Language Resources and Evaluation (LREC-2006)*, Genoa, Italy, 2006.

[11] K. Y. Itakura and C. L. Clarke. Using Dynamic Markov Compression to Detect Vandalism in the Wikipedia. In *SIGIR'09: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 822–823. ACM Press, 2009.

[12] S. Kullback and R. A. Leibler. On Information and Sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[13] D. McGuinness, H. Zeng, P. da Silva, L. Ding, D. Narayanan, and M. Bhaowal. Investigation into Trust for Collaborative Information Repositories: A Wikipedia Case Study. In *Proceedings of the Workshop on Models of Trust for the Web*, 2006.

[14] S. M. Mola-Velasco. Wikipedia Vandalism Detection Through Machine Learning: Feature Review and New Proposals. In M. Braschler and D. Harman, editors, *Notebook Papers of CLEF 2010 LABs and Workshops, 22-23 September, Padua, Italy*, Sept. 2010.

[15] M. Potthast. Crowdsourcing a Wikipedia Vandalism Corpus. In *Proc. of the 33rd Intl. ACM SIGIR Conf. (SIGIR 2010)*. ACM Press, Jul 2010.

[16] M. Potthast, B. Stein, and R. Gerling. Automatic Vandalism Detection in Wikipedia. In *ECIR'08: Proceedings of the 30th European Conference on IR Research*, volume 4956 of *LNCS*, pages 663–668. Springer-Verlag, 2008.

[17] M. Potthast, B. Stein, and T. Holfeld. Overview of the 1st International Competition on Wikipedia Vandalism Detection. In M. Braschler and D. Harman, editors, *Notebook Papers of CLEF 2010 LABs and Workshops, 22-23 September, Padua, Italy*, Sept. 2010.

[18] R. Priedhorsky, J. Chen, S. K. Lam, K. Panciera, L. Terveen, and J. Riedl. Creating, Destroying, and Restoring Value in Wikipedia. In *Group'07: Proceedings of the International Conference on Supporting Group Work*, 2007.

[19] E. J. Rodríguez Posada. AVBOT: detección y corrección de vandalismos en Wikipedia. *NovATIca*, (203):51–53, 2010.

[20] K. Smets, B. Goethals, and B. Verdonk. Automatic Vandalism Detection in Wikipedia: Towards a Machine Learning Approach. In *WikiAI'08: Proceedings of the Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy*, pages 43–48. AAAI Press, 2008.

[21] F. B. Viégas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with History Flow visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 575–582. ACM Press, 2004.

[22] A. G. West, S. Kannan, and I. Lee. Detecting Wikipedia Vandalism via Spatio-Temporal Analysis of Revision Metadata. In *EUROSEC'10: Proceedings of the Third European Workshop on System Security*, pages 22–28, 2010.

[23] Wikipedia contributors. Wikipedia: Vandalism — Wikipedia, The Free Encyclopedia, 2010. [accessed 23-Oct-2010].

[24] H. Zeng, M. Alhoussaini, L. Ding, R. Fikes, and D. McGuinness. Computing Trust from Revision History. In *Intl. Conf. on Privacy, Security and Trust*, 2006.