Detecting Wikipedia Vandalism using Machine Learning

Notebook for PAN at CLEF 2011

Cristian-Alexandru Drăguşanu, Marina Cufliuc, Adrian Iftene

UAIC: Faculty of Computer Science, "Alexandru Ioan Cuza" University,
General Berthelot, 16, 700483, Iasi, Romania
{cristian.dragusanu, marina.cufliuc, adiftene}@infoiasi.ro

Abstract. Wikipedia vandalism identification is a very complex issue, which is now mostly solved manually by volunteers. This paper presents the main components of a system built by our group in order to automatically identify vandalized Wikipedia articles. The main component of our system is a machine learning component that uses three types of features grouped in 3 classes: Metadata, Text and Language. Additional to previous approaches we consider 4 new features related to vulgar, biased, sexual and miscellaneous bad words. The obtained results showed an area of 0.42464 under the PR-AUC curve and an area of 0.82963 under the ROC-AUC curve.

1 Introduction

Wikipedia is the largest online encyclopedia. It is free to access by anyone and its main advantage is that it can also be edited by any user, at any time. This caused a rapid growth to its number of available articles and languages. At the moment of this writing, Wikipedia is available in 281 languages. Top 3 Wikipedias are, in order, English, German and French, each having over 1.000.000 articles. The English Wikipedia has over 3.600.000 articles constantly updated and maintained by over 140.000 active users and over 1.500 administrators.

The advantage of being a free encyclopedia which anyone can edit is also a significant problem, because, at any given time, any old or new article, in any language, is prone to being vandalized. PAN 2011[1] has a task called "Wikipedia Vandalism Detection", which targets the development of systems capable of detecting Wikipedia vandalism. According to the PAN 2010 Wikipedia Vandalism Detection training corpus [1], about 7% of all revisions were vandalized. This is a significant problem for Wikipedia, because the readers can never be sure of the quality of available information, unless they verify it from other sources. While some vandalism cases can be spotted very easily (such as improper language and massive text deletion), other times finding it is more difficult (such as fake information inserted in articles).

[1] PAN 2011: http://pan.webis.de/

Research studies in the field were made only in recent years and concluded that detection of vandalism is related to artificial intelligence. The best method, which is heading towards current research directions are focused on machine learning techniques [2] and the statistical analysis in natural language processing [3]. Also a good method of detection is based on spatial and temporal analysis of revisions made to the Wikipedia articles [4]. Other related articles treating automatic Wikipedia vandalism detection include [5], [6] and [7].

Since 2006 they created a series of automated tools to detect vandalism. These tools, called anti-vandalism bots, are programs that are designed to automatically detect and remove vandalism actions. What is the easiest method of disposal is to bring the document to the previous version identified by bots as act of vandalism.

Currently the most important bots are ClueBot[2] and VoABot II[3]. These tools use regular expressions and lists of database users or IP addresses blocked to prevent vandalism of articles. However these bots detect only about 30% of the total number of acts of vandalism, so it is necessary to improve methods of detection and correction of existing techniques.

The most notable results are currently achieved by combining the detection rules of STiki[4], Cluebot NG[5], WikiTrust[6] as well as an URL spam detection system.

In the following, we present the approach our group in an attempt to identify acts of vandalism in existing edits on Wikipedia. These edits were made available by the organizers of PAN 2011, part of CLEF 2011[7].

2 Edit Features and Classification

Our approach is based on the best performing detector at the time of this study [8] (according to the main Wikipedia Vandalism Detection page[8]). We removed some of the features and added a few others. All our features are grouped in 3 classes: *Metadata*, *Text* and *Language*. Our main target was to see how well a detector could work based solely on the information found in the training corpus, without using any additional information (such as external services like WikiTrust, or querying Wikipedia for detailed information about the author of the revisions or the history of the article). As a result, we didn't implement any reputation features (proposed in [8]), or features such as: TIME_SINCE_PAGE, TIME_SINCE_REG or TIME_SINCE_VAND. We did, however, try to use the Google SafeBrowsing service[9] to detect any possible malicious links that were inserted in new revisions. But this attempt was unsuccessful, because of two reasons:

[2] ClueBot: http://en.Wikipedia.org/wiki/User:ClueBot/Source

[3] VoABot: http://en.wikipedia.org/wiki/User:VoABot_II

[4] STiki: http://en.wikipedia.org/wiki/Wikipedia:STiki

[5] Cluebot NG: http://en.wikipedia.org/wiki/User:ClueBot_NG

[6] WikiTrust: http://en.wikipedia.org/wiki/WikiTrust

[7] CLEF 2011: http://clef2011.org/

[8] Wikipedia Vandalism Detection: http://www.uni-weimar.de/medien/webis/research/events/pan-11/wikipedia-vandalism-detection.html

[9] Google Safe Browsing API: http://code.google.com/apis/safebrowsing/

1. the training corpus didn't contain relevant information of this kind (there weren't sufficiently many cases in which vandalized revisions contained links marked by Google SafeBrowsing as malware/phishing);

2. the huge time difference between the date of the revisions, dated 2009, and the current Google SafeBrowsing results (there were a few cases where some URLs are currently considered dangerous, but 2 years ago they were OK). The same situation can be found while trying to use the Wikipedia URL blacklist[10], which now contains a few domains that, in the past, were perfectly OK.

So we didn't use the Google SafeBrowsing results in the final detection process. Of course, using such services for real-time, current revisions which take place on Wikipedia could provide very good results. But the use for detecting old vandalized revisions is very limited.

The complete list of used features follows below.

2.1 Features used by participants in PAN 2010

These features are explained in detail in [8].

Metadata features – generated based on general revision information:

- **IS_REGISTERED**: marks if the author of the edit has a Wikipedia account. This feature is not computed by querying Wikipedia for this information, but instead the editor name is checked to see if it represents a valid IP (anonymous edit) or not (registered user);
- **COMMENT_LENGTH**: the length of the edit revision;
- **SIZE_CHANGE**: length difference between the new and old revisions;
- **SIZE_RATIO**: ration between the new and old revisions text length;
- **PREV_SAME_AUTH**: if the old revision has the same author as the new one.

Text features – based on basic analysis on text characters:

- **DIGIT_RATIO**: the frequency of digits in the new revision;
- **ALPHANUM_RATIO**: the frequency of alpha-numeric characters in the new revision;
- **UPPER_RATIO**: the frequency of upper case characters in the new revision;
- **UPPER_LOWER_ RATIO**: ratio between the upper case and lower case characters in the new revision;
- **LONG_CHAR_SEQ**: longest single character sequence length;
- **LONG_WORD**: longest word length;

[10] Wikipedia Spam Blacklist: http://en.wikipedia.org/wiki/Wikipedia:Spam_blacklist

- **COMPRESS_LZW**: compression ratio of added words (using the LZW algorithm);
- **PREV_LENGTH**: the text length of the previous revision.

Language features – based on more advanced analysis over the text content; multiple word dictionaries were used to search the text for different words, belonging to different categories:

- **VULGARITY**: the frequency of vulgar words;
- **PRONOUNS**: the frequency of first and second person pronouns;
- **BIASED_WORDS**: the frequency of high bias words;
- **SEXUAL_WORDS**: the frequency of non-vulgar sexual words;
- **MISC_BAD_WORDS**: the frequency of any other words with negative meaning (or not suitable for an encyclopedia);
- **ALL_BAD_WORDS**: the frequency of all bad words (vulgar, pronouns, biased, sexual and miscellaneous);
- **GOOD_WORDS**: the frequency of words that are not bad;
- **COMM_REVERT**: if the new revision comment marks that previous changes were reverted to an earlier state.

2.2 Customized Features

We customized a few features from [9], [10] and used them in the **Language** class: VULGARITY2, BIASED_WORDS2, SEXUAL_WORDS2 and MISC_BAD_WORDS2. Their description is presented below:

- **VULGARITY2**: the ratio between the frequency of vulgar words in the new revision and their frequency in the old revision;
- **BIASED_WORDS2**: the ratio between the frequency of high bias words in the new revision and their frequency in the old revision;
- **SEXUAL_WORDS2**: the ratio between the frequency of non-vulgar sexual words in the new revision and their frequency in the old revision;
- **MISC_BAD_WORDS2**: the ratio between the frequency of miscellaneous bad words in the new revision and their frequency in the old revision.

The purpose of these features is to distinguish articles which use the words from the targeted categories in a legitimate way (vulgar, biased, sexual or miscellaneous bad words). For instance, there might be non-vandalized articles which already have a high frequency of words from the above categories. Inevitably, any new revisions to those articles will still have a high frequency for those words, in which case, new revisions might have features which resemble those of a vandalism, even though the revisions might not be vandalism. Examples of such articles would be the articles titled Profanity[11], Seven Dirty Words[12] and other.

Basically, if a previous revision (considered non-vandalized) contains a high frequency of words from the above categories, then it might be normal that new revisions have a similar high frequency for those word categories. And the features we added attempt to mark these special situations, by comparing the frequencies in the old and new revisions. These features are meant to treat a few special cases that were not correctly treated by the features from section 2.1.

2.3 Classifier

After all features have been computed for the training corpus, a classifier model has been trained using a Support Vector Machine algorithm. We used the LibSVM library[13], using the C-Support Vector Classification SVM type and Radial Basis Function (RBF) kernel type [11]. All features were scaled in the [0, 2] interval and the SVM algorithm has been set to train a model which can also output probability estimates, which made it possible to show exact confidence values.

3 Evaluation

We submitted one run to the PAN 2011 at Wikipedia Vandalism Detection task for English language. The run was obtained using LibSVM with the features presented above. Computing the features took around 9 hours for all training revisions and about 24 hours for the test corpus. After all features were computed, training the SVM model and classifying the test revisions was done a lot faster, in under 1 hour.

Our tests also showed that most detection problems we had were with blanked revisions. There were two situations when this occurred. Firstly, in cases when a vandalism occurred by blanking an article, which lead to the new revision being blank. And secondly, when such vandalism was reverted, in which case the old revision was blank and the new one wasn't.

In both cases, the SVM algorithm had problems classifying the revisions correctly, because the revision features had either very low values (0), or very high (infinite, in cases where ratios were computed and the denominator was a feature which was 0). We attempted to correct to some degree these situations by applying a few post-classification rules and treat specifically the blank revisions classification, by lowering (when a revision was reverted) or increasing (when the new revision was blank) their final confidence level.

[11] Profanity: http://en.wikipedia.org/wiki/Profanity

[12] Seven Dirty Words: http://en.wikipedia.org/wiki/Seven_Dirty_Words

[13] LibSVM library: http://www.csie.ntu.edu.tw/~cjlin/libsvm/

3.1 Official results

The official results[14] published by the organizers are presented in Table 1 and in Figures 1, 2. The results were obtained using PR-AUC and ROC-AUC measures presented in [12].

Table 1: Results of UAIC's runs

English Wikipedia Vandalism			
Rank	**PR-AUC**	**ROC-AUC**	**Participant**
1	0.82230	0.95313	A.G. West, University of Pennsylvania, USA
2	0.42464	0.82963	A. Iftene and C.-A. Dragusanu, AL.I.Cuza University, Romania

From [12] we have that plotting precision versus recall spans the precision-recall space, and plotting the *TP* (the number of edits that are correctly identified as vandalism, i.e. *true positives*) rate versus the *FP* (the number of edits that are untruly identified as vandalism, i.e. false positives) rate spans the ROC space.

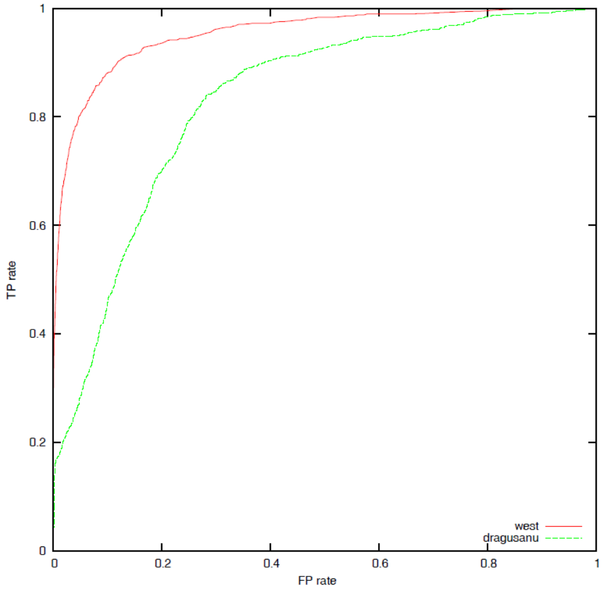

Fig. 1. Evaluation of submitted runs to Wikipedia Vandalism Detection task using ROC measure

[14] Evaluation Results: http://pan.webis.de/

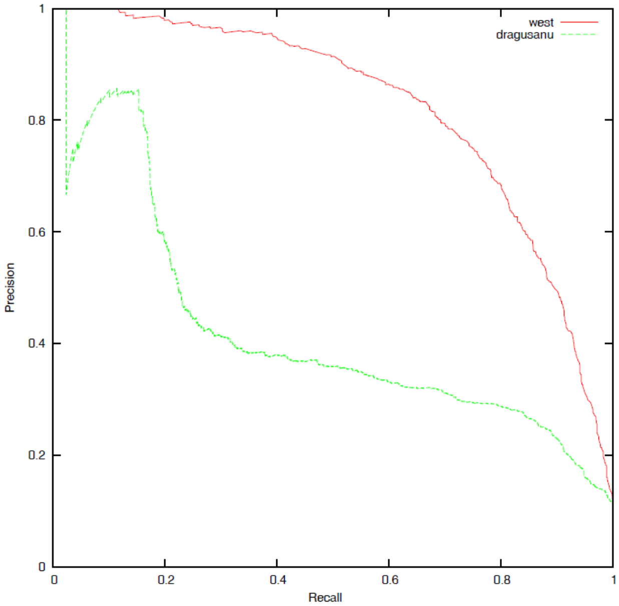

Fig. 2. Evaluation of submitted runs to Wikipedia Vandalism Detection task using precision-recall-curve (PR-AUC)

From Table 1 we can see how the results of A.G. West group are better than our results. According to the PR measure, their result is much better (see Figure 2), and according to the ROC measure the results are closer (see Figure 1).

4 Conclusions

In this paper we presented our group's participation in the PAN 2011 exercise in Wikipedia Vandalism Detection task from CLEF 2011 labs.

In the future we also intend to use a more advanced natural language processing method (for instance, to extract and compare the main ideas from the old revision and the new revision) because we believe that this area can bring significantly improved results to our system. Natural language processing is the closest way of interpreting the actual meaning of the text in the same manner as the human brain does, and so determining the real meaning of the words could offer valuable information for detecting article vandalism.

Acknowledgements. The research presented in this paper was funded by the Sector Operational Program for Human Resources Development through the project "Development of the innovation capacity and increasing of the research impact through post-doctoral programs" POSDRU/89/1.5/S/49944.

References

1. Potthast, M.: Crowd sourcing a Wikipedia Vandalism Corpus. 33rd Annual International ACM SIGIR Conference (SIGIR 10), Geneva, Switzerland, ISBN 978-1-4503-0153-4 (2010)
2. Smets, K., Goethals, B., Verdonk, B.: Automatic Vandalism Detection in Wikipedia: Towards a Machine Learning Approach. Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI) Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy (WikiAI08) (2008)
3. Chin, S. C., Street, W. N.: Detecting Wikipedia vandalism with active learning and statistical language models. WICOW'10, North Carolina, USA. (2010)
4. West, A. G., Kannan, S., Lee, I.: Detecting Wikipedia Vandalism via Spatio-Temporal Analysis of Revision Metadata. Technical Reports (CIS), University of Pensylvania, Department of Computer & Information Science. (2010)
5. Priedhorsky, R., Chen, J., Lam, S. K., Panciera, K., Terveen, L., Riedl, J.: Creating, Destroying, and Restoring Value in Wikipedia. Group'07: Proceedings of the International Conference on Supporting Group Work, Sanibel Island, Florida, USA (2007)
6. Itakura, K. Y., Clarke, C. L. A.: Using Dynamic Markov Compression to Detect Vandalism in the Wikipedia. SIGIR'09: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA (2009)
7. Geiger, R. S., Ribes, D.: The Work of Sustaining Order in Wikipedia: The Banning of a Vandal. CSCW'10: Proceedings of the ACM Conference on Computer Supported Cooperative Work, Savannah, Georgia, USA, (2010)
8. Adler, B., de Alfaro, L., Mola-Velasco, S. M., Rosso, P., West, A.: Wikipedia Vandalism Detection: Combining Natural Language, Metadata, and Reputation Features. Computational Linguistics and Intelligent Text Processing, University of California, Santa Cruz, USA (2011)
9. Potthast, M., Stein, B., Gerling, R.: Automatic Vandalism Detection in Wikipedia. Advances in Information Retrieval: Proceedings of the 30th European Conference on IR Research (ECIR 2008), Glasgow, UK, 4956 of Lecture Notes in Computer Science, Springer. ISBN 978-3-540-78645-0 (2008)
10. Mola-Velasco, S. M.: Wikipedia Vandalism Detection Through Machine Learning: Feature Review and New Proposals - Lab Report for PAN at CLEF 2010, Notebook Papers of CLEF 2010 LABs and Workshops, Padua, Italy, ISBN 978-88-904810-0-0 (2010)
11. Chang, C.-C., Lin, C.-J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, New York, NY, USA (2011)
12. Potthast, M., Stein, B., Holfeld, T.: Overview of the 1st International Competition on Wikipedia Vandalism Detection, In CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy, ISBN 978-88-904810-0-0 (2010)