PressAcademia Procedia - (PAP), ISSN: 2459-0762

PressAcademia Procedia

2nd World Conference on Technology, Innovation and Entrepreneurship
May 12- 14, 2017, Istanbul, Turkey. Edited by Sefer Şener

USING TEXTUAL FEATURES FOR THE DETECTION OF VANDALISM IN WIKIPEDIA: A COMPARATIVE APPROACH IN LOW-RESOURCE LANGUAGE SECTIONS

DOI: 10.17261/Pressacademia.2017.575
PAP-WCTIE-V.5-2017(13)-p.80-87

Arsim Susuri[1], Mentor Hamiti[2], Agni Dika[3]
[1]South East European University, Faculty of Contemporary Sciences and Technologies , arsimsusuri@gmail.com
[2]South East European University, Faculty of Contemporary Sciences and Technologies , m.hamiti@seeu.edu.mk
[3]South East European University, Faculty of Contemporary Sciences and Technologies , agnidika@yahoo.com

ABSTRACT

This study investigates the impact of using textual features for the detection of vandalism across low-resource language sections in Wikipedia. For this purpose, we propose new features that allow the machine learning-based text classifiers to better distinguish vandalism and to improve the detection rates of vandalism across languages, based on textual features applied in previous researches. These features enable us to compare the contributions of the bots against vandalism, stressing the differences between bots and editors with regards to the detection of vandalism. We propose a new set of efficient and language independent features, which has the performance level similar to the previous sets. Three Wikipedia sections will be used for this purpose: Simple English (simple), Albanian (sq) and Bosnian (bs). We will show that our set of textual features has similar and, in some cases, better vandalism detection rates across languages than previous research.

Keywords: Wikipedia, textual features, low-resource languages, vandalism.

1. INTRODUCTION

Vandalism is a great challenge for Wikipedia, with humans being the main cause, through various illegitimate acts leave traces in computer systems. Our hypothesis is that vandalism can be characterized through models of article views of vandalized Wikipedia articles and that vandalism behaviour is similar across different languages. In the past, a similar research was done by West (2013) and Tran and Christen (2013). This paper is an extension of work originally presented in Susuri et al. (2016), by addressing the issue of using textual features detecting vandalism in Wikipedia's articles across languages. According to our hypothesis, a model developed in one language can be applied to other languages. If successful, this would drop the costs of training the classifiers separately for each language. Applying this model of vandalism detection across different languages shows similar results. In this paper, we will explore the possibility of applying the detection of vandalism across languages through textual features. We combine these data sets in order to analyse any gains in terms of language independency of certain features.

For this purpose, we compare performances of standard classifiers for identifying vandalism in three Wikipedia data sets (Simple English, Albanian and Bosnian). On top of this, we compare the performances of classifiers in one language and the other one and in the combined, data set. Vandalism detection bots have and are changing the way Wikipedia identifies and prevents vandals (Geiger and Ribes, 2010). However, contributions from these bots are rarely discussed, despite their importance for maintaining the reliability of Wikipedia, some of whom have become the most active editors (Adler et al., 2008). The increase of responsibilities of bots in the detection of vandalism raises several research issues:

- How big is the difference in values between the bots of vandalism detection and editors?
- How do they differ across different languages?

Our research is based on learning to detect vandalism bots together as well as from the editors, and then implemented these models in three different languages: English (simple), Albanian (sq) and Bosnian (bs).

We propose a new set of efficient and language independent features, with similar or better performance levels in comparison to previous textual features. We will show that bots and users have similar reference values in terms of identifying vandalism when their models apply to other sets of classification in cases of vandalism. The contributions of this paper are:

- Development of new textual features used for the detection of vandalism, independent of language, with better efficiency in comparison to previous researches
- Determining the differences between bots and users regarding vandalism identified by bots and users
- Demonstration of application of classification models across languages without losing out on quality classification

The paper is organized as follows. Section 2 describes data sets used for this purpose, and the textual features used for the purpose of the detection of vandalism across languages. Section 3 reports and discusses the empirical results. And the final section gives conclusions.

2. DATA SETS AND FEATURES

2.1. Data Sets

Wikipedia saves all article revisions in the so-called history dumps in XML or SQL format. For the purpose of this paper will use the history dumps of Wikipedia in three low-resource language section: history dumps in Albanian[1], Simple English[2] and in Bosnian[3]. We use the Wikipedia history dumps of edits dated 29.10.2015, for Simple English, Albanian and Bosnian language sections. History files (*pages-meta-history*) include article revision history, including additional data related to the revision and labeling text.

The reason for choosing these data sets is that, to our knowledge, no one else has analysed the history dump of Wikipedia in Albanian in the context of detection of vandalism. Two other history dumps were selected to validate our hypothesis about the independence of the textual features, to validate detection of vandalism across languages, and because the sizes of these data set are similar. In terms of the volume of articles, similarities exist between these three editions of Wikipedia, as described below:

- Wikipedia in Simple English contains 119,183 articles[4]
- Wikipedia in Albanian contains 55,894 articles[5]
- Wikipedia in Bosnian contains 69,765 articles[6]

Therefore, we can compare the possibility of applying our textual features in the detection of vandalism. We split these data sets into training sets (revisions before 2015) and testing sets (revisions in 2015). The testing sets contain 8-28% of all revisions for each of the three language sections. We will distinguish contributions of bots from contributions of editors (users), and compare vandalism that one classifier can learn from vandalism repairs made from bots or users.

2.2. Feature Creation

We create our textual features stemming from differences in the content of the repaired revision of the article in comparison to the previous revision, containing vandalism. Using the *diff* algorithm (Hunt and McIlroy, 1974) we obtain unique rows in terms of revision before the repair, unique rows in terms of revision after the repair and revised rows during the process of repair. We do not take into account common rows, due to precise designation of changes in the content. Common rows show us the ratio of vandalized content and the legitimate content, but in cases of mass deletions, the size of unique rows in relation to the repaired revision is sufficient to indicate vandalism.

The features are shown in Table 1.

We then apply the process of determining the difference at the word level with the aim of extracting vandalized words that were repaired. For the processing of the text, we use Unicode (UTF-8) and alphabets for the English, Albanian and Bosnian language, respectively. Along with the appropriate description and the average time for generating features, we also use the statistical test of Kolmogorov-Smirnov (K-S) (Massey, 1951). Our features are applied on changed words instead of changed revisions (as in previous researches). We use textual features from PAN 2010 and PAN 2011 Workshops (Mola-Velasco, 2010; West and Lee, 2011), where we find the first application of textual features for the purposes of detecting vandalism.

[1] https://dumps.wikimedia.org/sqwiki/
[2] https://dumps.wikimedia.org/simplewiki/
[3] https://dumps.wikimedia.org/bswiki/
[4] https://simple.wikipedia.org/wiki/Main_Page
[5] https://sq.wikipedia.org/wiki/Faqja_kryesore
[6] https://bs.wikipedia.org/wiki/Početna_strana

Features from V01-NR1 to V10-NTF are created from the revisions before and after the repair. Features from V11-FP to V18-KV are created from changed words during the repair process, which isolate possible vandal words and include distribution of words in the repair. Features from V19-RMV to V24-SGV are applied on each of the repaired words, where values indicating vandalism are chosen.

2.2.1. Modifying Features

The above-mentioned features are suitable for detecting changes in the content of the articles.

Features from V01-N1 to V04-NRN2 are enumerations of types of rows from the *diff* algorithm. High values of unique rows in the vandalized revision (before repair) could be a sign of massive insertions, while higher values in the repaired revision (after repair) could be a sign of massive deletions. The number of rows changes may be a sign of small changes as a sign of vandalizing insertions or textual changes.

Table 1: List of Features Created Before (1) and After (2) Repair

Features	Description	Time (ms)	K-S test (failures)	K-S test (PAN) (failures)
V01-NR1	Number of Unique Rows in (1)	0.043	10%	0%
V02-NR2	Number of Unique Rows in (2)	0.043	0%	50%
V03-NRN1	Number of Unique Revised Rows in (1)	0.043	10%	50%
V04-NRN2	Number of Unique Revised Rows in (2)	0.043	10%	25%
V05-DGTF	Difference in Total Length of Unique Words in (1) and (2)	0.522	0%	25%
V06-DNTF	Difference in Total Number of Unique Words in (1) and (2)	0.472	0%	0%
V07-RGTF	Ratio of Total Length of Unique Words in (1) and (2)	0.522	11%	25%
V08-RNTF	Ratio of Total Number of Unique Words in (1) and (2)	0.472	11%	25%
V09-NFV	Number of Unique Words	0.005	12%	0%
V10-NTF	Number of Total Words	0.003	125%	0%
V11-FP	Pronoun Words	0.011	50%	100%
V12-FZ	Slang Words	0.007	30%	50%
V13-FV	Vulgar Words	0.007	50%	100%
V14-FSM	Capitalised Words	0.006	10%	0%
V15-FAN	Alphanumerical Words	0.006	10%	0%
V16-SV	Single Letters	0.007	80%	100%
V17-NN	Single Digits	0.005	22%	75%
V18-KV	Single Characters	0.006	80%	100%
V19-RMV	Highest Ratio Uppercase-Lowercase	0.170	0%	25%
V20-RSS	Highest Ratio Digits-All Letters	0.170	0%	25%
V21-GFP	Length of the Most Repeated Word	0.200	12%	50%
V22-GKG	Length of the Longest Character Repeated	0.175	10%	50%
V23-FGV	Longest Unique Word	0.045	10%	25%
V24-SGFV	Sum of Length of the Unique Words	0.045	10%	0%

These features are similar to the feature byte change in the West and Lee (2011), but the difference here is the possibility of clarifying the impact of changes in these features. Features from V05-DGTF to V10-NTF are similar to the enumeration of rows. Through these features the process of enumerating changes in words, before and after repair, is implemented. These changes indicate cases of minor vandalism that modify specific words. The difference in length of specific words, and the difference in numbers of specific words determine the relative size changes needed for revision repair. Lengths and numbers of special words determine the relative change in size and the sheer number of changes needed to repair the vandalism. These combinations of features can identify repairs made by bots in cases of minor vandalism.

2.2.2. Features from Wiktionary

We borrow features from Wiktionary[7] in English[8], Albanian[9] and Bosnian[10] languages. At the same time, we are borrowing the workshops features of PAN Workshops, adapted for our testing purposes. By using features from V11-FP to V13-FV we analyse the ratio letter / word considering three types of actual vandalism: pronouns, slang and vulgarism. In total, there are 40 pronouns, 1130 slangs and 347 and vulgarisms. We seek these words in the *diff* algorithm at the sentence level. For example, if vulgarism in the English language is used in the Albanian language, these vulgarism features are numerated into

[7] https://www.wiktionary.org/
[8] https://simple.wiktionary.org/wiki/Main_Page
[9] https://sq.wiktionary.org/wiki/Faqja_kryesore
[10] https://bs.wiktionary.org/wiki/Početna_strana

the features for revision in the Albanian language. Visual inspection shows that slang and vulgarism are not usually common words in other languages. Features from V14-FSM to V18-KV are used to enumerate different types of sentences. By analysing the letters of each word, some indications of possible vandalism can include words that start with uppercase, words with numbers and words that consist of only one letter. These features are common indicators of vandalism in other studies (Mola-Velasco, 2010; West and Lee, 2011), as well.

2.2.3. Features from Words

Features presented in Table 1 are based on preliminary research but have been modified in order to apply in our analysis at the level of words. In analysing the change of sentence, the idea is that an unusual appearance a presentation of unusual words point to vandalism, therefore, not apply a total or average value based on the fact that the vandaliser can attempt to avoid detection of vandalism by disguising the vandalism with legitimate question but not related to sentences vandalized. Features from V19-RMV to V20-RSS are used to determine the ratio letter/word. These features are based on research conducted by Mola-Velasco (2010), with a difference in that the application of our experiments is at the level of words and not at the level of the document. Minimum or maximum values are a clear indicator of vandalism. Features from V21-GFP to V22-GKG are used to determine the length of the longest repeated character in a word, as applied in Mola-Velasco (2010), as one of the clear indicators of vandalism. Feature from V23-FGV to V24-SGFV are used to separate longest unique words and total size of unique words in the changed words. These features are based on Mola-Velasco (2010) and West and Lee (2011), but with a modified application.

3. EXPERIMENTS

We apply separation of the Wikipedia data sets into training sets (all reviews prior to October 2015) and into testing sets (after October 2015). Since the data set is not balanced, we apply under-sampling in order to balance the legitimate revisions with vandalised revisions. Through this preparatory procedure we enable the Random Forest algorithm to improve performance with more balanced sample trees.

3.1. Classification Results

We use the Random Forest classifier in Weka[11]. This classifier is better in terms of performance, as shown in previous relevant research (Adler et al., 2011), therefore this is the reason of the application of this classifier in our tests. To get the most out of the performance of this classifier, we apply a ten-fold cross validation on the training data, using a number of parameters of this classifier, such as the number of the forecasters (the trees in the forest), the maximum number of features, the minimum number of leaf samples, the minimum number of samples for separation and minimal density. The results obtained are measuring values of Area Under Curve – Precision Recall (AUC-PR). The reason for using AUC-PR is that these values obtained are not affected by the design of the data (under-sampling in our case) (Davis and Goadrich, 2006).

3.1.1 Combining Classification Languages

For complete sets of Wikipedia (simple and sq), Table 2 presents the results of the combination of training and testing data. Within the same language and type of users (diagonal values), the classifier shows higher performance in comparison with language combinations. Except in the case of the bots' values in Albanian, where the classifier trained on the data from bots in Wikipedia in English shows better results. This suggests that the bots of Wikipedia in English can identify more cases of vandalism in the Albanian section of Wikipedia than bots of Albanian section of Wikipedia.

Table 2: Classification Across Languages and Users for the Random Forest Classifier

AUC-PR	**Testing**	**simple**		**sq**		**bs**	
Training	***Type***	Bots	users	bots	users	bots	users
simple	Bots	0.923	0.821	0.871	0.732	0.913	0.778
	Users	0.912	0.844	0.849	0.755	0.905	0.785
sq	Bots	0.899	0.766	0.864	0.730	0.854	0.764
	Users	0.893	0.785	0.847	0.749	0.834	0.783
bs	Bots	0.912	0.801	0.881	0.787	0.918	0.768
	Users	0.905	0.823	0.847	0.802	0.907	0.790

For users in three languages, we find consistent performance in the detection of vandalism in three languages. This suggests that users search for similar patterns of vandalism just as bits do. Users of Simple English Wikipedia identify, proportionally, more cases of vandalism across languages than users of other languages. This suggests that with more users (editors), more vandalism models can be identified.

[11] http://www.cs.waikato.ac.nz/ml/weka/

3.1.2 Combined Training Data

We combine training data from bots and users for both languages, for each type of editor, and for two languages and two types kinds of editors. The classification results are presented in Table 3. The purpose of this test is to try to apply the learning of vandalism without distinguishing the contribution from bots or users. By learning from the bots and users for each language, we find some differences in the classification performance. Bots follow common rules and structures of vandalism that learning algorithms learn fast, providing more accurate results and higher AUC-PR value. On the other hand, cases of vandalism detection from editors have greater variation in vandalism types as there are cases of vandalism which cannot be detected by bots. Similarly, as in the previous case, we find no statistically significant difference when comparing the rows between Tables 2 and 3. This shows that there are no differences in learning vandalism from bots or users across languages.

So, the combination of observations from bots and users does not cause performance improvement in the vandalism detection rates identified by users. It makes us realize that, on one hand, users really identify a wider array of vandalism types, while on the other hand, the contributions of bots do not change with the changing of languages, but nevertheless improve the classification performance. Although these improvements are small, in reality it means hundreds or thousands of cases (depending on the section of Wikipedia) of vandalism are automatically detected.

Table 3: Classification Across Languages and Users

AUC-PR	***Testing***	**simple**		**sq**		**bs**	
Training	***Type***	botët	përdoruesit	botët	përdoruesit	botët	përdoruesit
simple	Bots/users	0.932	0.808	0.902	0.732	0.911	0.785
sq	Bots/users	0.924	0.798	0.905	0.735	0.896	0.782
bs	Bots/users	0.927	0.803	0.915	0.734	0.934	0.784
Total	Bots	0.937	0.793	0.931	0.730	0.931	0.783
Totai	Users	0.927	0.815	0.953	0.741	0.937	0.791
Total	Bots/users	0.935	0.816	0.947	0.754	0.934	0.795

3.1.3. Combined Training and Testing Data Sets

To complete the learning across languages and to have comparable data with previous research, we combine both types of editors for training and testing data. In Table 4 we have presented the results of the classification across languages for each language and combined training for the world and for three languages user total data and training. Results of the relevant training and testing data of the languages in question (Simple English, Albanian and Bosnian) show AUC-PR values being between the values in Table 3 and 4. This enables us to understand that by using all training data, detection rates benefit significantly better statistically in both languages.

Table 4: AUC-PR values across three languages

AUC-PR	**Testing**		
Training	simple	sq	bs
Simple	0.874	0.734	0.804
Sq	0.865	0.732	0.811
Bs	0.872	0.722	0.833
Bots	0.871	0.723	0.802
Users	0.871	0.734	0.754
Total	0.884	0.743	0.834

3.1.5 Results of Data Modelling

We have modeled over-fitted legitimate reviews by the Random Forest classifier because it enables us to build decision-making tree on the classifier in order to distinguish vandalism, reduces the size of the model and data needed for training, and reduces the time of the learning. However, modeling the data raises questions of bias in performance. Table 4 shows values according to the ratio 1:1 of legitimate reviews with vandalised reviews (over-fitted values). In order to have more convincing results, we repeated our experiments based on modelled ratios of 2:1, 5:1, 9:1 and 13:1. The latter two values are applied in the experiments because the share of the vandaised reviews in the PAN-WVC-10 and PAN-WVC-11 data sets is approximately 7% (93% reviews legitimate - 7% vandalised reviews) which coincides with the ratio 13: 1. Ratio 9: 1 reflects the range of vandalism found in previous studies (Potthast, 2010).

We compare the values obtained in Figure 1, for the classification within the same language (diagonal values in Table 4). For classification across languages, in Figure 2 we have shown average PR-AUC values, along with mean standard error. Based

on figures and applied tests, we can conclude that the re-modelling process has little negative effet on the values of classification. In Figure 1 we have shown the results of the training applied to the balanced data,with a ratio of 1: 1, together with the results of applying on unbalanced testing data sets with ratios 8:1 and 13:1. These results simulate real consequences of learning in the balanced data set and applying in the unbaanced data set, such as full sections Wikipedia. In Figure 3 we have shown the results of the classification within the same language, as well as in Figure 4 the classification across languages, with average AUC-PR values, along with mean standard error. Experiments applied in PAN data sets are included in the Figures 3 and 4 for comparative purposes.We can see from the figures, that the proposed features have very good results, compared with previously applied features. Similarly, for the classification across languages, the proposed features have very good and comparable values to previous research. On this basis, we can say that there is a stable trend of vandalism within a language, which is applicable to a large extent, across languages.

4. FINDINGS AND DISUSSIONS

The benefit from the application of machine learning across languages for the detection of vandalism is the generalization of classification models for different sections of Wikipedia, without learning from many cases of vandalism. Our results show that learning the language that has many cases of vandalism, such as the section on

Figure 1: Comparison of Different Values of Classification across Languages

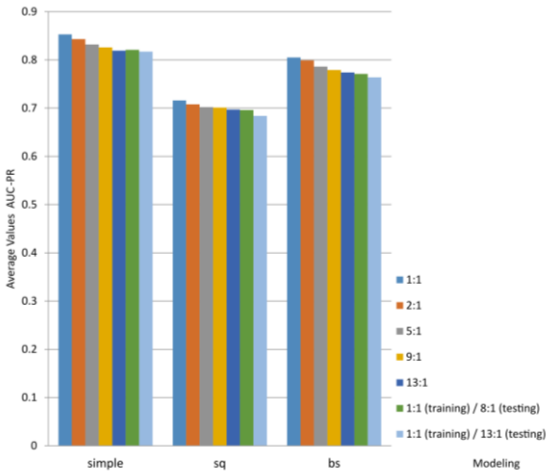


Figure 2: Comparison of Different Features of Classification within Languages

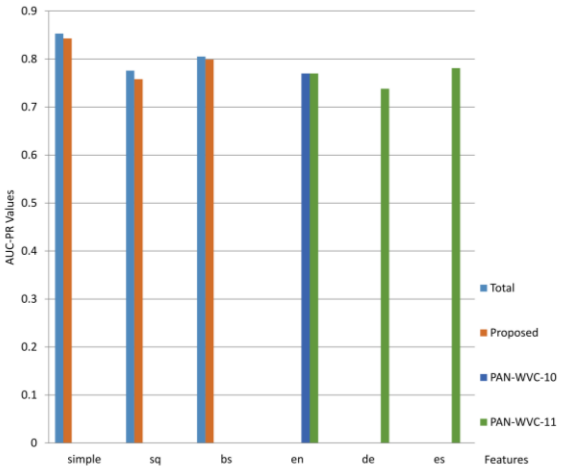
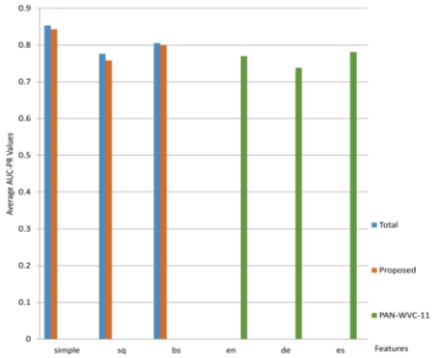
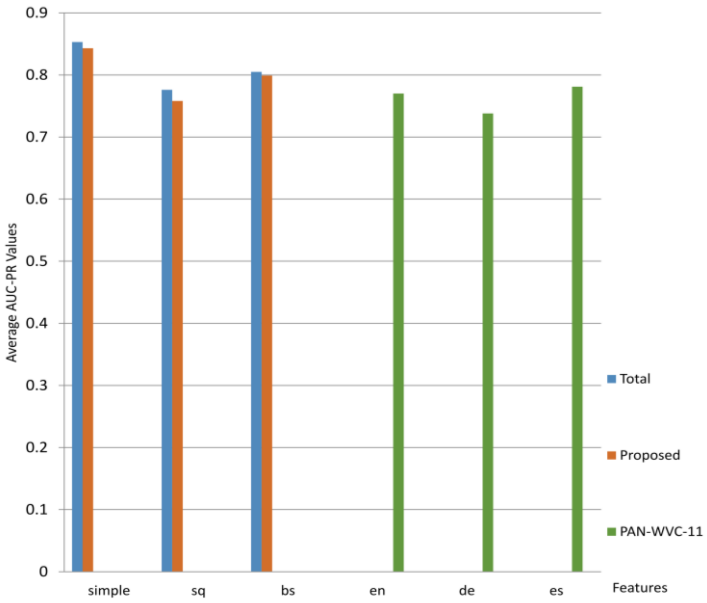


Figure 3: Comparison of Combination of Different Features of Classification across Languages



Simple English may generalize very well to smaller sections of Wikipedia, such as the sections in Albanian and in Bosnian. One advantage of our approach is the immediate and comparative analysis of the current revision of the text of the previous review, with the aim of determining the potential vandalism.

Figure 4: Comparison of Combinations of Different Features of Classification across Languages



We do not need to add meta data, and user profiling to determine vandalism.

Our textual features show performance which is comparable to the previous research and improves performance based on the preliminary review models in Wikipedia. Our features are designed with the aim of generalizing the language in experimentation, which is reflected in the classification performance.

A limitation of our research is to support the textual features that cannot easily detect vandalism which can detect whether we use metadata or user features reputation. Our classification method uses under-fitting as a balanced approach to distinguish between the majority class (legitimate reviews) and minority class (vandalizing revisions), and to reduce the training data set. Although this method can add noise in the data sets, results in Section 3.1.5 tell us that under-fitting does not affect the classification results significantly, if we repeat the experiments with different proportions of training and testing. We have presented only one classifier performance which, although in these sections Wikipedia has shown consistent results, may not be the best in other sections of Wikipedia.

5. CONCLUSION

In this chapter we applied the comparison between bots and users, in terms of identifying based on the detection of vandalism in low-resources sections of Wikipedia. We have developed textual features that include features commonly used for the detection of vandalism and we used a classifier for the features, listed above, important to bots and to users in

three language sections of Wikipedia. We present and discuss the differences between the bots and users in terms of identifying vandalism in three sections of Wikipedia: Simple English, Albanian and Bosnian languages. Comparison with previous research has shown that our techniques are comparable and sometimes better than these researches. Our contributions showed that we can apply the learning of vandalism in one language section of Wikipedia and then apply a classifier in other sections with little loss in quality of classification.

In the future, the focus of the research should be anonymous users' contribution in identification of vandalism. This is because of difficulties in determining their identity. Creating an online system for the purpose of evaluating the efficiency of the characteristics previously submitted and to assess the performance impact of these features in the detection of vandalism, will further improve the detection of vandalism.

REFERENCES

Adler B. T, de Alfaro L., Pye I., 2008, "Measuring author contributions to the Wikipedia. In: WikiSym '08, Porto, Portugal, 8-10 September 2008. New York: ACM.

Adler B. T., de Alfaro L., Mola-Velasco S. M., Rosso P., and West A. G., 2011, "Wikipedia vandalism detection: Combining natural language, metadata, and reputation features". In Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part II, CICLing'11, pages 277 - 288, Berlin, Heidelberg, Springer-Verlag.

Davis J. and Goadrich M., 2006, "The Relationship Between Precision-Recall and ROCCurves". In Proceedings of the 23rd International Conference on Machine learning (ICML), 2006.

Geiger R. S. and Ribes D., 2010, "The Work of Sustaining Order in Wikipedia: The Banning of a Vandal". In Proceedings of the 22nd ACM Conference on Computer Supported Cooperative Work (CSCW).

Hunt J. W., Mcllroy M. D, 1974, "An Algorithm for Differential File Comparison", Computer Science Technical Report, Bell Laboratories.

Massey F. J., 1951, "The Kolmogorov-Smirnov Test for Goodness of Fit". Journal of the American Statistical Association, 46.

Mola-Velasco S. M., 2010, "Wikipedia Vandalism Detection Through Machine Learning: Feature Review and New Proposals". In CLEF (Notebook Papers/Labs/-Workshops).

Susuri A., Hamiti M. and Dika A, 2016, "Machine Learning Based Detection of Vandalism in Wikipedia across Languages". In proceedings of the 5th Mediterranean Conference on Embedded Computing (MECO), Bar, Montenegro.

Tran K.N., Christen P., 2013 "Cross-language prediction of vandalism on wikipedia using article views and revisions". Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD).

West A. G., 2013, "Damage Detection and Mitigation in Open Collaboration Applications", Ph.D. thesis, University of Pennsylvania.

West A. G. and Lee I., 2011, "Multilingual Vandalism Detection using Language-Independent & Ex Post Facto Evidence". In CLEF (Notebook Papers/Labs/Workshops).