Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)

Spam Users Identification in Wikipedia via Editing Behavior

Thomas Green, Francesca Spezzano
Computer Science Department
Boise State University
Boise, Idaho 83702
thomasgreen@u.boisestate.edu, francescaspezzano@boisestate.edu

Abstract

In this paper, we address the problem of identifying spam users on Wikipedia and present our preliminary results. We formulate the problem as a binary classification task and propose a set of features based on user editing behavior to separate spammers from benign users. We tested our system on a new dataset we built consisting of 4.2K (half spam and half benign) users and 75.6K edits. Experimental results show that our approach reaches 80.8% classification accuracy and 0.88 mean average precision. We compared against ORES, the most recent tool developed by Wikimedia which assigns a damaging score to each edit, and we show that our system outperforms ORES in spam users detection. Moreover, by combining our features with ORES, classification accuracy increases to 82.1%. Additionally, we also show that our system performs well in a more realistic, unbalanced setting, i.e. when spammers are greatly outnumbered by benign users, by achieving an AUROC of 0.84 (which increases to 0.86 when we combine with ORES).

Introduction

Social media connects millions of people across the globe. These extensive networks provide users a way to connect and share information with others regardless of spatial proximity to one another. This ultra-connectivity also attracts the sharing of malicious content, e.g. spam, due to the large potential audience on social media.

Wikipedia, the online, user-maintained encyclopedia, for example, is viewed by hundreds of millions of users every month. Given its large audience and open format of allowing users to edit pages, Wikipedia is a major target for users posting malicious content. Broadly speaking, Wikipedia declares "any malicious edit which attempts to reverse the main goal of the project of Wikipedia"[1] as vandalism. Within this broad definition, various categories of vandalism do exist: standard vandalism is the creation (or deletion) of content to damage the integrity of information on a page, trolling is the creation of content with the main purpose of upsetting others and creating a hostile environment, and spamming is the unsolicited promotion of some entity. Specifically, Wikipedia recognizes three main types of spam, namely "advertisements masquerading as articles, external link spamming, and adding references with the aim of promoting the author or the work being referenced"[2].

Past work addressed the problem of detecting damaging edits (especially vandalism (Adler et al. 2011)) by looking at edit content through linguistic features and URL properties (West et al. 2011a). Various tools are currently running on Wikipedia to detect vandalism (Cluebot_ NG ; STiki) or damaging edits in general (ORES), but nothing specific for spam detection. Even with these bots working, however, detection mechanisms are still not perfect, and spammers still manage to post spam messages with varying levels of success. The majority of the work to protect Wikipedia from spammers is done manually by Wikipedia users (patrollers, watchlisters, and readers) who monitor recent changes in the encyclopedia and, eventually, report suspicious spam users to administrators for definitive account blocking.

In this paper we address the problem of Wikipedia spamming from a different perspective and study the problem of identifying spam users, instead of spam edits. We make the following contributions. (1) We propose a machine learning-based framework using a set of features which are language independent and based on user editing behavior to identify spam users. (2) To test our framework, we built new dataset containing 4.2K (half spam and half benign) users and 75.6K edits. (3) We experimentally show that our system is able to classify spammers from benign users with 80.8% of accuracy and significantly improves over past work. Moreover, we show that our system is valuable in suggesting potential spammers to Wikipedia administrators for further investigation as proved by a mean average precision of 0.88. Finally, we show that, even in the more realistic case where we have more benign users than spammers, our system performs pretty good with an AUROC of 0.84. By combining our system with ORES, the most recent tool developed by Wikimedia to assign a damaging score to each edit, our performances improve.

Related Work

Various efforts have been made in the past to detect spammers on social networks, mainly by studying their behavior

[1] http://en.wikipedia.org/wiki/Wikipedia:Vandalism

[2] http://en.wikipedia.org/wiki/Wikipedia:Spam

after collecting their profiles through deployed social honeypots (Stringhini, Kruegel, and Vigna 2010; Lee, Caverlee, and Webb 2010). Generally, social networks properties (Song, Lee, and Kim 2011; Yang, Harkreader, and Gu 2011), posts content (Grier et al. 2010), and sentiment analysis (Hu et al. 2014) have been used for spammer detection.

On Wikipedia, plenty of work has been done on detecting damaging edits, particularly vandalism (see (Adler et al. 2011) for a survey). Our work is closer in spirit to (Kumar, Spezzano, and Subrahmanian 2015) in which a behavior-based approach is used to discern vandal users from benign ones. Regarding spam detection specifically, previous work concentrated on the problem of predicting whether a link contained in an edit is spam or not, whereas, in this paper, we predict whether a user is a spammer or not by considering her edit history. (West et al. 2011b) created the first Wikipedia link-spam corpus, identified Wikipedia's link spam vulnerabilities, and proposed mitigation strategies based on explicit edit approval, refinement of account privileges, and detecting potential spam edits through a machine learning framework. The latter strategy, described by the same authors in (West et al. 2011a), relies on features based on (i) article metadata and link/URL properties, (ii) HTML landing site analysis, and (iii) third-party services used to discern spam landing sites. This tool was implemented as part of STiki (STiki), a tool which suggests potential vandalism edits to humans for definitive classification, and has been used on Wikipedia since 2011. Nowadays, this STiki component is inactive due to a monetary cost for third-party services, therefore we do not compare with this tool in our paper. Beyond STiki, another bot, ClueBot NG (Cluebot_ NG), also runs on Wikipedia and is used to detect vandalism. Recently, Wikimedia launched a new service, called ORES (Objective Revision Evaluation Service) (ORES), which measures the level of general damage (provided through a "damaging score") each edit causes. Since ClueBot NG and STiki are tailored toward vandalism detection, we compare our tool with ORES only.

Our approach looks at editing behavior instead of edit content and has the advantage of being general and applicable not only for English, but also for different other language versions of Wikipedia. Moreover, we do not rely on third-party services, so there is no overhead cost.

Dataset

This section describes the dataset containing both spam and benign users we collected through the publicly available Wikipedia API. We considered several lists provided by English Wikipedia as ground truth of users' usage status.

Initially, we collected all Wikipedia users (up to Nov. 17, 2016) who were blocked for spamming from two lists maintained on Wikipedia: "Wikipedians who are indefinitely blocked for spamming" [3] and "Wikipedians who are indefinitely blocked for link spamming" [4]. The first list contains all spam users blocked before Mar 12, 2009, while the second one includes all link-spammers after Mar 12, 2009 to today [5]. We gathered a total of 2,087 spammers (we only included users who did at least one edit) between the two lists considered.

In order to create a balanced dataset of spam/benign users, we randomly select a sample of benign Wikipedia users of roughly the same size as the spam user set (2,119 users). To ensure these were genuine users, we cross-checked their usernames against the entire list of blocked users provided by Wikipedia [6]. This list consists of all Wikipedia users who have been blocked for any reason, spammers included.

For each user in our dataset, we collected up to their 500 most recent edits. For each edit we gathered the following information: edit content, time-stamp, whether or not the edit is done on a Talk page, and the damaging score provided by ORES. Our final dataset consists of a total of 4.2K (half spam and half benign) users and 75.6K edits. The dataset is available at (SpamDataset).

Features for Spammers Identification

The features we propose are based on typical behaviors exhibited by spammers: similarity in edit size and links used in revisions, similar time-sensitive behavior in edits, social involvement of a user in the community through contribution to Wikipedia's Talk page system, and chosen username. We did not consider any feature related to edit content so that our system would be language independent and capable of working for all Wikipedia versions. Also, the duration of a user's edit history, from the first edit to her most recent edit, is not taken into account as this feature is biased towards spammers who are short-lived due to being blocked by administrators.

The list of features considered in our system are described in the following.

User's edit size based features.

Average size of edits - since spammers in Wikipedia are primarily trying to promote themselves (or some organization) and/or attract users to click on various links, the sizes of spammers' edits are likely to exhibit some similarity when compared to that of benign users.

Standard deviation of edit sizes - since many spammers make revisions with similar content, the variation in a user's edit sizes is likely not to be very large when compared to benign users.

Variance significance - since variance in a spam user's edits can change based on a user's average edit size, normalizing a user's standard deviation of edit sizes by their average edit size may balance any difference found by considering the standard deviation alone.

Editing time behavior based features.

Average time between edits - spammers across other social media tend to perform edits in batches and in relatively rapid succession, while benign Wikipedia users dedicate more time in curating the article content and then make edits more slowly than spammers.

[3] http://bit.ly/blocked_for_spamming

[4] http://bit.ly/blocked_for_link_spamming

[5] These two lists are not available anymore. We provide them at (SpamDataset).

[6] http://bit.ly/Block_List

	Accuracy	Precision (Benign Users)	Precision (Spammers)	Recall (Benign Users)	Recall (Spammers)	MAP
Our Features						
K-Nearest Neighbor	0.711	0.734	0.691	0.668	0.755	0.733
SVM	0.670	0.640	0.720	**0.790**	0.549	0.746
Logistic Regression	0.792	0.806	0.778	0.773	0.812	0.838
Random Forest	0.805	0.835	0.779	0.764	0.847	0.856
XGBoost	**0.808**	**0.839**	**0.781**	0.764	**0.851**	**0.880**
ORES	0.697	0.759	0.658	0.584	0.812	0.695
Our Features + ORES	**0.821**	**0.845**	**0.800**	**0.789**	**0.853**	**0.886**

Table 1: Performances of our features and comparison with ORES according to accuracy, precision, recall, and Mean Average Precision (MAP) metrics. ORES and Our Features + ORES are computed with XGBoost.

Standard deviation of time between edits - the consistency in timing of spammers' edits tends to be somewhat mechanical, while benign users tend to edit more sporadically.

Links in edit based features.

Unique link ratio - since spammers often post the same links in multiple edits, a measure of how unique any links that a user posts may be very useful in helping to determine which users are spammers. This measure is calculated for any user that has posted a minimum of two links in all of their edits, and it is the ratio of unique links posted by a user to the total number of links posted by the user (considering only the domain of the links)

Link ratio in edits - since spammers on Wikipedia are known to post links in an effort to attract traffic to other sites the number of edits that a user makes which contain links is likely a useful measure in determining spammers from benign users.

Talk page edit ratio. Since talk pages do not face the public and are only presented to a user that specifically clicks on one, spammers are less likely to get very many views on these pages, and, therefore are much less likely to make edits to talk pages. Because of this, the ratio of talk pages edited by a user that correspond with the main article pages that a user edits is considered a possible good indicator of whether a user is a spammer or not.

Username based features. (Zafarani and Liu 2015) showed that aspects of users' usernames themselves contain information that is useful in detecting malicious users. Thus, in addition to the features based on users' edit behaviors, we also considered four additional features related to the user's username itself. These four features are: the *number of digits in a username*, the *ratio of digits in a username*, the *number of leading digits in a username*, and the *unique character ratio in a username*.

Implementation and Experiments

In order to test the features we are proposing for the classification task, we considered different classifiers, namely Support Vector Machine (SVM)[7], Logistic Regression, K-Nearest Neighbor, Random Forest, and XGBoost. To evaluate the performances, we considered accuracy, precision, and recall metrics, and performed 10-fold cross validation. Results are reported in Table 1. XGBoost and Random Forest performed the best, with XGBoost having slightly higher values and reaching an accuracy of 80.8%. However, precision for the class of spammers in below 80%, causing potential blocking of several benign users. Nevertheless, when our tool is used to suggest potential spammers to Wikipedia administrators for further investigation, we obtain a good Mean Average Precision (MAP) of 0.88.

Feature analysis. To analyze our features, we computed feature importance via a forest of randomized trees. The relative importance (for the classification task) of a feature f in a set of features is given by the depth of f when it is used as a decision node in a tree. Features used at the top of the tree contribute to the final prediction decision of a larger fraction of the input samples. The expected fraction of the samples they contribute to can thus be used as an estimate of the relative importance of the features. Figure 1 shows the importance of our set of features for the classification task. The red bars in the plot show the feature importance using the whole forest, while the blue bars represent the variability across the trees. The top three most important features are: *Link ratio in edits*, *Average size of edits*, and *Standard deviation of time between edits*. As expected, spammers use more links in their edits. The average value of this feature is 0.49 for spammers and 0.251 for benign users. Also, benign users put more diverse links in their revisions than spammers (0.64 vs. 0.44 on average). We also have that spammer's edit size is smaller and they edit faster than benign users. Regarding edits on talk pages, we have that the majority of the users are not using talk pages (percentage for both benign users and spammers is 69.7%). However, surprisingly, we have that, among users editing talk pages, the talk page edit ratio is higher for spammers (0.2) than for benign users (0.081) and we observe a group of around 303 spammers trying to gain visibility by making numerous edits on talk pages.

Finally, username based features contribute to an increase in accuracy prediction by 2.9% (from 77.9% to 80.8%) and Mean Average Precision by 0.019 (from 0.861 to 0.880).

Comparison with ORES. The Objective Revision Evaluation Service (ORES) is a web service developed by Wikimedia Foundation that provides a machine learning-based scoring system for edits. More specifically, given an edit, ORES is providing three probabilities predicting (i) whether

[7]We used LibSVM library. The best performing SVM was *nu*-SVC with sigmoid kernel.

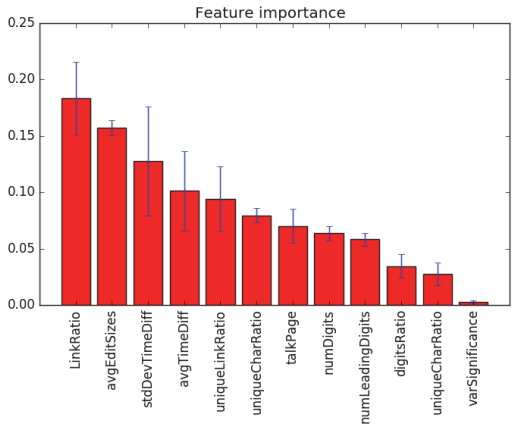

Figure 1: Feature importance.

or not an edit causes damage, (ii) if it was saved in good-faith, and (iii) if the edit will eventually be reverted. These scores are available through the ORES public API [8].

To compare our system with ORES, we considered the edit damaging score. More specifically, given a user and all her edits, we computed both the average and maximum damaging score provided by ORES and used these as features for classification. Results on 10-fold cross validation with XGBoost (the best performing classifier) are reported in Table 1. As we can see, ORES performances are poor for the task of spammer detection (69.7% of accuracy and a mean average precision of 0.695). However, combining our features with ORES helped to increase the accuracy to 82.1%. All other metrics also showed improvement except recall for spam users and MAP, which remained the same.

Unbalanced Setting. In reality, spam users are greatly outnumbered by benign users. Thus, we created an unbalanced dataset to test our method by randomly selecting users at a ratio of 10% spammers and 90% of benign users (due to the size of the data we have, we could not reduce this ratio further). Then, we performed 10-fold cross validation and measured the performance by using the area under the ROC curve (AUROC). To deal with class imbalance, we oversampled the minority class in each training set by using SMOTE (Chawla et al. 2002). We also considered class weighting, but we found that SMOTE is performing the best. Due to the randomness introduced, we repeated each experiment 10 times and averaged the results.

Table 2 reports the results for this experiment. As we can see, even with class imbalance, our features reach a good AUROC of 0.842 (in comparison we have an AUROC of 0.891 for the balanced setting) and significantly improve over ORES (AUROC of 0.736). Adding ORES features to ours helps to increase the AUROC to 0.864.

[8] http://ores.wikimedia.org

	AUROC
ORES	0.736
Our Features	**0.842**
Our Features + ORES	**0.864**

Table 2: Our features vs. ORES performance in the unbalanced setting. Everything is computed by using XGBoost.

Conclusions

In this paper we presented our preliminary research on the problem of identifying spam users on Wikipedia. We showed that our behavior-based approach achieves an 80.8% classification accuracy and 0.88 mean average precision, outperforms past work, and works well in an unbalanced setting (AUROC of 0.842). As we did not use any linguistic features on edit content, our system can work on different language versions of Wikipedia.

References

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 *CICLing*, 277–288.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16:321–357.

Cluebot_NG. http://bit.ly/ClueBotNG.

Grier, C.; Thomas, K.; Paxson, V.; and Zhang, M. 2010. @ spam: the underground on 140 characters or less. In *CCS*, 27–37.

Hu, X.; Tang, J.; Gao, H.; and Liu, H. 2014. Social spammer detection with sentiment information. In *ICDM*, 180–189.

Kumar, S.; Spezzano, F.; and Subrahmanian, V. 2015. Vews: A wikipedia vandal early warning system. In *KDD*, 607–616.

Lee, K.; Caverlee, J.; and Webb, S. 2010. Uncovering social spammers: Social honeypots + machine learning. In *SIGIR*, 435–442.

ORES. http://bit.ly/wikipedia_ores.

Song, J.; Lee, S.; and Kim, J. 2011. Spam filtering in twitter using sender-receiver relationship. In *RAID*, 301–317.

SpamDataset. http://bit.ly/wiki_spammers.

STiki. http://bit.ly/STiki_tool.

Stringhini, G.; Kruegel, C.; and Vigna, G. 2010. Detecting spammers on social networks. In *ACSAC*, 1–9.

West, A. G.; Agrawal, A.; Baker, P.; Exline, B.; and Lee, I. 2011a. Autonomous link spam detection in purely collaborative environments. In *WikiSym*, 91–100.

West, A. G.; Chang, J.; Venkatasubramanian, K.; Sokolsky, O.; and Lee, I. 2011b. Link spamming wikipedia for profit. In *CEAS*, 152–161.

Yang, C.; Harkreader, R. C.; and Gu, G. 2011. Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers. In *RAID*, 318–337.

Zafarani, R., and Liu, H. 2015. 10 bits of surprise: Detecting malicious users with minimum information. In *CIKM*, 423–431.