Cross Language Learning from Bots and Users to detect Vandalism on Wikipedia

Khoi-Nguyen Tran, *Student Member, IEEE,* and Peter Christen

Abstract—Vandalism, the malicious modification of articles, is a serious problem for open access encyclopedias such as Wikipedia. The use of counter-vandalism bots is changing the way Wikipedia identifies and bans vandals, but their contributions are often not considered nor discussed. In this paper, we propose novel text features capturing the invariants of vandalism across five languages to learn and compare the contributions of bots and users in the task of identifying vandalism. We construct computationally efficient features that highlight the contributions of bots and users, and generalize across languages. We evaluate our proposed features through classification performance on revisions of five Wikipedia languages, totaling over 500 million revisions of over 9 million articles. As a comparison, we evaluate these features on the small PAN Wikipedia vandalism data sets, used by previous research, which contain approximately 62,000 revisions. We show differences in the performance of our features on the PAN and the full Wikipedia data set. With the appropriate text features, vandalism bots can be effective across different languages while learning from only one language. Our ultimate aim is to build the next generation of vandalism detection bots based on machine learning approaches that can work effectively across many languages.

Index Terms—Bots, cross language learning, editors, feature engineering, transfer learning, users, vandalism, Wikipedia

1 Introduction

The prevalence of Wikipedia as the largest free and open access online encyclopedia attracts millions of volunteer contributors and tens of millions of article views every day [1]. As a result, Wikipedia attracts many types of vandals that deliberately make malicious edits. Each edit to Wikipedia is recorded as a revision, where the latest revision of an article is displayed to readers. Cases of vandalism are seen in the revision history of many articles across many languages. To combat vandalism, editors can repair the damage or revert the latest revision to a previous revision, where they usually leave a comment to indicate the occurrence of vandalism. Wikipedia distinguishes many types of vandalism, which are generally in one of the categories defined by Priedhorsky et al. [2]: "misinformation, mass delete, partial delete, offensive, spam, nonsense, and other", where "other" means (possibly new) types of vandalism behavior not covered by any defined categories.

Vandalism is often caught and repaired quickly [2], [3], [4], but the number of cases of vandalism grows in proportion to the fast growth of Wikipedia. Our large data sets (discussed in Section 3) totalling over 500 million revisions of over 9 million articles show editors identified an average of over 2,100 cases of vandalism per day in 2012 for the English Wikipedia. To identify and repair this many cases each day, automated vandalism detection programs – known as bots – have been developed to partially relieve the burden on editors. Through keyword search of edit comments, bots (bot editors - 0.67%) and users (human editors - 1.33%) repair vandalism in nearly 2% of all revisions in the English Wikipedia [3]. This contrasts with other studies – using crowdsourced votes from manual inspection of a sampled set of revisions – showing vandalism may appear in 7% to 11% of all revisions [5]. These missing cases of vandalism (approximately 5% to 9%) suggest very difficult or ambiguous forms of vandalism that may require up to 8 rounds of majority consensus from three different annotators in each round [5].

The use of counter-vandalism bots is changing the way Wikipedia identifies and bans vandals [6], [7]. However, contributions by bots are often not considered nor discussed, despite their importance to Wikipedia and some bots becoming the most prolific editors [6], [8]. The increasing delegation of vandalism detection to bots poses interesting research questions: how do the detection rates of bots and users compare to each other, and how do they differ across different Wikipedia languages?

In this paper, we investigate these questions by learning vandalism collectively recognized by bots and users, and evaluating these models against both bots and users across 500 million revisions from five different languages: English (en), German (de), Spanish (es), French (fr), and Russian (ru). We propose a new set of computationally efficient features that are language invariant, and have classification performance comparable to the previously proposed features. We show bots and users have similar vandalism identification scores when we apply them on the other's recognized set of vandalism cases. Fur-

• *K.-N. Tran and P. Christen are with the Research School of Computer Science, The Australian National University, Canberra, Australia. E-mail: khoi-nguyen.tran@anu.edu.au, peter.christen@anu.edu.au*

thermore, we show that combinations of vandalism classification models generalize well across languages without statistically significant loss in classification quality. To strengthen our results, we replicate our experiments on the baseline vandalism data sets of approximately 62,000 revisions from competitions held for the PAN Workshops [5], and discuss limitations with these data sets.

Our contributions are (1) developing novel text features that capture language invariant aspects of vandalism, and have greater effectiveness compared to features from related work as demonstrated by a statistical test and feature ranking; (2) contrasting the differences between bots and users by learning vandalism identified by bots and users; (3) demonstrating that cross language application of classification models do not have significant loss in classification quality; (4) conducting our experiments on the entire Wikipedia data dumps (over 500 million revisions), which comprehensively includes all random samples of revisions in the PAN baseline data sets; and (5) replicating our experiments on these much smaller baseline data sets, showing and contrasting the performance of features often used in related work on these data sets and on the full Wikipedia data dumps.

The rest of this paper is organized as follows. Section 2 reviews the related work. Section 3 describes the Wikipedia data sets we use in our work, and Section 4 details and ranks the language invariant text features of vandalism. Section 5 describes our cross language learning method, and Section 6 summarizes and compares our results to the PAN data sets. Section 7 compares our results to related work, and Section 8 discusses our findings, advantages, and limitations. Finally, Section 9 provides conclusions and future directions of research.

2 Related Work

We begin by discussing bots and editing applications used for vandalism detection to show the problem of detecting vandalism in the larger context of the Wikipedia community, and then summarize vandalism research by the type of data sets.

Bots and editing applications. Bots are an integral part of Wikipedia because they provide automation to repetitive and mundane tasks, but their contributions are often ignored in research or by the Wikipedia community [9]. For example, activities of some bots do not appear on the list of recent changes provided by Wikipedia [9]. The prolific editing activity of bots and their discreetness have lead to mistrust by some editors because the perceived aggressiveness of bots in completing their task without regards to the social dynamics of the editing communities surrounding each article [9]. Interruptions of bots in tasks such as detecting vandalism can greatly increase exposure and longevity of vandalism, but they also show the resilience of Wikipedia to eventually restore order [10]. The importance of bots to Wikipedia is seen through their editing contributions and their influence on the editing culture of Wikipedia through interactions with users across many languages of Wikipedia [7].

Counter-vandalism bots and counter-vandalism applications also suffer backlash from users (see Section 8), which could be attributed to incorrect identification of vandalism. These bots – in particular, ClueBot[1] and ClueBot NG[2] – are evolving to use machine learning techniques to detect more sophisticated forms of vandalism, which takes time to learn correct cases of vandalism. The counter-vandalism applications that come with user-interfaces are changing their design to guide editors in identifying and softer handling of potential vandalism cases from incoming edits. Some examples are Huggle[3], one of the most popular; STiki [11], developed from research on user reputation for vandalism detection; and Snuggle [12], developed through research on user interface design and socialization of bots on Wikipedia.

Sampling Wikipedia and small Wikipedias. Vandalism detection research is often performed on samples of the English Wikipedia. A featureless compression method for detecting vandalism is presented by Itakura and Clarke [13] on randomly selected articles. Words can be predictors of whether an article will be reverted as demonstrated by Rzeszotarski and Kittur [14]. Revisions made by bots are analyzed, but evaluation and comparison of classification performance is only for revisions of one Wikipedia article.

The cross language application of classification models is a type of transfer learning [15]. Chin et al. [16] apply transfer learning to detect vandalism on Wikipedia by learning vandalism from one article and applying the models to another article. Revisions from the Webis Wikipedia vandalism corpus [17] are segmented and placed into similar clusters. The best performing vandalism classification models built on each cluster are then evaluated on clusters from revisions of two selected English Wikipedia articles.

Some Wikipedias do not need bots as they are small and have sufficient human editors to manage all articles. Smets et al. [18] use the Simple English Wikipedia (499,395 revisions of 53,449 articles) to evaluate vandalism techniques based on bag-of-words and Naive Bayes, and Probabilistic Sequence Modeling. The classifiers are compared to the performance of two rule-based counter-vandalism bots. Arguments for the need of machine learning for the vandalism detection task are presented in the paper.

PAN Workshop Data Sets. The interpretation of vandalism differs amongst Wikipedia users, which can lead to incomplete or inconsistent labeling of vandalized revisions on Wikipedia. Potthast et al. [5]

1. http://en.wikipedia.org/wiki/User:ClueBot
2. http://en.wikipedia.org/wiki/User:ClueBot_NG
3. http://en.wikipedia.org/wiki/Wikipedia:Huggle

develop two corpora by crowd-sourcing votes on whether a Wikipedia revision contains vandalism using Amazon's Mechanical Turk. The corpus PAN-WVC-10 contains around 32,000 revisions sampled from the English Wikipedia, where 7% of the revisions contain vandalism. The corpus PAN-WVC-11 contains less than 10,000 revisions for each of the English, German, and Spanish Wikipedias, where approximately 11% of all revisions contain vandalism.

The PAN Workshops in 2010 and 2011 held competitions to build machine learning based vandalism detectors from these corpora. For the PAN-WVC-10 data set, Velasco [19] uses a set of 21 features to detect vandalism, which resulted in a first place ranking at the 2010 workshop. Adler et al. [20] improve on the winning entry of the PAN 2010 Workshop by adding metadata, text, user reputation, and language features, totaling 37 features. These features are evaluated individually and in combinations using a Random Forest classifier, where using all features show the best performance. Similarly, Javanmardi et al. [21] further improve the classification results by introducing 66 features and applying feature reduction. Combinations of features are also explored to determine the best feature sets to detect vandalism.

Other techniques showing improvements to the winner of the 2010 PAN Workshop focus on analyzing other properties of the revision content for vandalism [11], [22], [23], [24], [25], [26]. The main drawback of these other techniques is that they are not scalable because of the deep text and structure analysis that are costly in time to generate features when applied to the entire Wikipedia data.

For the PAN-WVC-11 data sets, West and Lee [27] develop 65 features that include many of the features from the entries for the 2010 PAN Workshop. These features are described generally as language independent, ex post facto (developed after recognition of vandalism), and language driven features. A classifier built on these features resulted in a first place ranking at the 2011 PAN Workshop for each language [27]. However, classification in non-English Wikipedia revisions showed very poor performance in the AUC-PR scores (0.708 for German, and 0.489 for Spanish) compared to English Wikipedia revisions (0.822), but comparable performance in the AUC-ROC scores (0.969 for German, 0.868 for Spanish, and 0.953 for English).

We use the PAN Workshop data sets in our research as a baseline comparison. The samples in the data sets do not have many revisions made by bots to learn from. The PAN-WVC-10 data set contains 14 bots with a total of 101 revisions (0.3%), where one bot is a counter-vandalism bot that made a total of 25 revisions (0.07%). The PAN-WVC-11 data set contains a total of 7 bots across three languages, with a total of 34 revisions (0.1%), where one bot is a counter-vandalism bot that made a total of 5 revisions (0.02%). Clearly, we cannot effectively learn and compare bots and users with these few revisions made by bots.

Using all Wikipedia revisions. Features extracted from the metadata of revisions allow all Wikipedia article revisions to be processed because of their relative simplicity compared to the revision content. West et al. [28] explore a variety of features generated from the metadata of all Wikipedia article revisions for detecting vandalism. The reputation features on article, user, category, and country show interesting variations and sources of vandalism.

The Wikipedia article views data set[4] is understudied because of its size and linking required with the revision content. Our past research [1] uses all Wikipedia article revisions and views to detect vandalism in the English and German Wikipedias. We compare five classifiers and do not observe significant loss in classification quality when applying models across languages. We use metadata features derived from two Wikipedia data sets: revisions and views, where the latter has not been used for vandalism detection. In this work, we focus on developing content text features that show the contributions of bots and users across five languages, and leave the inclusion of metadata features as future work.

3 Wikipedia Data Sets

Wikipedia provides monthly data dumps of every language edition. We downloaded the first data dump available in 2013 and use all revisions from 2001 to December 31st 2012 (our cut off date) for these five languages: English (en), German (de), French (fr), Spanish (es), and Russian (ru). We chose these languages because they have some of the highest number of articles on Wikipedia, where four are the United Nations official languages and the most spoken languages in the world. We can provide our data parsing scripts and data sets on request.

3.1 Data Processing

The Wikipedia data dumps contain revisions for every article, but we only use the encyclopedic articles (namespace 0) as these articles are the reason people access Wikipedia. Every edit made on an article on Wikipedia generates a new revision with the full content of the article. When vandalism is discovered, it is usually repaired by correcting the vandalized content or by reverting to a past revision, which copies the past revision to become the current revision. In either case, the repaired revision may contain keywords – such as "rvv" (revert due to vandalism), "vandalism", "...rv...vandal...", and analogues in the other languages – in its comment indicating vandalism was detected and repaired.

4. http://dumps.wikimedia.org/other/pagecounts-raw/

TABLE 1
Number of unique editors (bots and users) in our data sets. An active editor is one that has made an edit in December 2012.

Editor	Bots		Users	
Wiki	Total	Active (%)	Total	Active (%)
en	925	121 (13.1%)	31,427,529	438,629 (1.4%)
de	876	81 (9.3%)	6,347,974	63,960 (1.0%)
es	443	80 (18.1%)	5,030,842	82,330 (1.6%)
fr	478	85 (17.8%)	3,557,384	60,115 (1.7%)
ru	323	88 (27.2%)	2,138,513	63,649 (3.0%)

As we are only interested in the textual features derived from the revision content, we reduce data size by focusing on the difference in the content of the flagged revision with the previous revision. We use the Python unified diff[5] algorithm to obtain lines (marked by a full stop or period) unique to each revision and the lines changed.

To distinguish revisions made by bot editors, we obtain lists of bot names for each language from Wikipedia articles and categories maintaining these lists[6]. We split the revisions into those made by bots and those made by users. We do not distinguish edits made by counter-vandalism tools, nor anonymous and registered users, which we leave as future work.

Using this data processing method, we found approximately 1.6% of all revisions from the English encyclopedic articles are identified cases of vandalism, which is consistent with the method and results from Kittur et al. [3]. Our work focuses on vandalism that triggers a bot or user to repair the revision. We are not interested in all vandalism cases because from visual inspection of some revisions we find that vandalism is sometimes missed and not usually expanded on, which leads to successive revisions containing the same or very similar vandalism. This will likely result in higher classification scores as the true positive class contains repeated samples. Our rationale is to find revisions that trigger counter-vandalism bots and users to interpret as vandalism, and not the successive revisions containing vandalism that may not have been inspected by counter-vandalism bots and users.

3.2 Data Statistics

Table 1 provides a count of the number of bots and users as found in our data sets. In total, we found 2,053 unique bots amongst all bots reported across the five languages. Wikipedia defines an active user as one having performed an action in the last 30 days, which we interpret in our data sets as a user having performed an edit in December 2012. Our visual inspection of bot names shows many bots have worked or are working across different languages, where some have not reported to or have not been identified by that language community on Wikipedia. We also find many bots are reported as active on Wikipedia, but have not made a contribution to any encyclopedic Wikipedia article. Counter-vandalism bots identify the majority of vandalism, but many other bots have some contributions to detecting and repairing vandalism.

Table 2 summarizes the number of revisions in our data sets split by editor type and revision type. For learning (see Section 5), we further split the data sets into training sets (all revisions before 2012) and testing sets (all revisions in 2012). The testing sets contain between 9-30% of all revisions for each language.

5. http://docs.python.org/2/library/difflib.html
6. E.g. https://en.wikipedia.org/wiki/Wikipedia:Bots/Status

TABLE 2
Number of article revisions in different languages, split by revision type, and bots and users.

Wiki	Type	Regular		Caught Vandals	
	Editor	Bots	Users	Bots	Users
en	Count	23,577,853	293,243,092	1,819,782	3,592,394
	(%)	7.4%	92.6%	33.6%	66.4%
	Total	316,820,945 (98.4%)		5,115,045 (1.6%)	
de	Count	8,274,593	60,564,993	4,754	189,551
	(%)	12.0%	88.0%	2.5%	97.5%
	Total	68,839,586 (99.7%)		194,305 (0.3%)	
es	Count	8,956,251	32,870,538	218,748	128,189
	(%)	21.4%	78.6%	63.1%	36.9%
	Total	41,826,789 (99.2%)		346,937 (0.8%)	
fr	Count	12,885,088	42,524,023	48,101	169,888
	(%)	23.3%	76.7%	22.1%	77.9%
	Total	55,409,111 (99.6%)		217,989 (0.4%)	
ru	Count	6,710,919	26,192,505	182	46,978
	(%)	20.4%	79.6%	0.4%	99.6%
	Total	32,903,424 (99.9%)		47,160 (0.1%)	
PAN 2010 en	Count	100	29,945	1	2393
	(%)	0.3%	99.7%	0.1%	99.9%
	Total	30,045 (92.6%)		2,394 (7.4%)	
PAN 2011 en	Count	24	8,818	0	1,143
	(%)	0.3%	99.7%	0%	100%
	Total	8,842 (88.5%)		1,143 (11.5%)	
PAN 2011 de	Count	6	9,395	0	589
	(%)	0.1%	99.9%	0%	100%
	Total	9,401 (94.1%)		589 (5.9%)	
PAN 2011 es	Count	4	8,889	0	1081
	(%)	0.1%	99.9%	0%	100%
	Total	8,893 (89.2%)		1,081 (10.8%)	

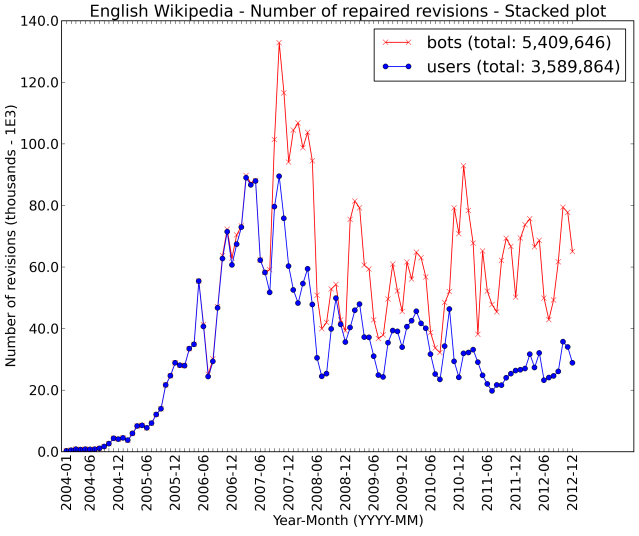

Fig. 1
Stacked line plot of the number of vandalized revisions identified by bot and users each month in the English Wikipedia.

We show the increasing use of bots to detect vandalism each month in the English Wikipedia in Figure 1. In the other Wikipedia languages, we do not see this trend because there may be a bias towards developing bots for the English Wikipedia, a mistrust of bots, or a smaller number of articles for each editor to maintain.

Overall, the regular revisions show bots are actively working in other languages and namespaces of Wikipedia with activity similar to users working on regular revisions. We see bots sharing a large portion of the workload of over 7%, but with vandalism detection, there is significantly lower usage of bots in non-English Wikipedias. Nevertheless, bots are an important resource for Wikipedia across its languages, and their contributions to vandalism detection cannot be ignored or neglected.

4 Feature Engineering

We generate our features from words extracted from the difference of the content of the repaired revision with the previous revision, which contains vandalism. From the diff algorithm, we have lines (separated by periods) unique to the revision before the repair, lines unique to the revision after the repair, and the lines changed in the repairing process. We ignore common lines to accurately determine changes in content. The common lines can show the ratio of the vandalized content to normal content, but for cases such as mass deletes, the size of lines unique to the repaired revision is sufficient to show this case. We further perform a sentence difference to extract vandal words that were repaired. Our text processing uses unicode (UTF-8) encoding and language specific alphabets.

All features are shown in Table 3 with a summarized description, an average time of generating features in milliseconds (ms), and a Kolmogorov-Smirnov (K-S) statistical test [29] (described in Section 4.4). We order our features in groups of relatedness, where bolded features are our novel contributions to detecting vandalism. Note that our features are applied specifically to diff words instead of the full diff of revisions as in previous works. Our borrowed features are text features from the winners of the PAN 2010 and PAN 2011 Workshops [19], [27], where they first appeared for the use of detecting vandalism.

Features `F00` to `F09` are generated from the revisions before and/or after a repair. Features `F10` to `F20` are generated from the words changed in the repair, which isolate possible vandal words and captures distributions of words in the repair. Note that duplicate words can exist and we count these in some features. Features `F21` to `F31` are applied on each word that was repaired, where we select for values that indicate vandalism. Although some features are derivatives from related work, we justify their novelty by our application to lists of single words – further polarizing vandalism cases – and show their effectiveness on the full Wikipedia data set.

4.1 Data Modification Features

Although these features are novel, they are intuitive in capturing the changes in content. We focus on changes reported by our diff algorithm.

Features `F00` to `F03`: These features are a count of types of lines from the diff algorithm. High counts of unique lines in the vandalized revision (before the repair) indicate mass insertions, and high counts in the repaired revision (after repair) indicate mass deletions. The count of line changes indicates small changes that may show vandalized insertions or changes of text.

Features `F04` to `F09`: Similar to the line counts, we count the changes of words before and after a repair. These changes in the words of the repair show the subtler cases of vandalism that modify specific words. The difference of word lengths and number of words show the extreme changes needed to repair vandalism, whereas the ratios show the relative size of changes needed for repair. Similarly, the lengths and the counts of the unique words show the relative change in size and the absolute number of changes needed in repairing vandalism. These combinations ensure that we can identify some of the repairs made by bots and users of subtler vandalism.

4.2 PAN Workshop Features

We borrow these features directly from the winners of the PAN workshops, where they have been often used by related work (see Section 2). The features are adapted for our data sets where needed and we provide clearer sources for vulgar and slang words.

Features `F10` to `F12`: Three types of words common or indicative of vandalism are pronouns, slang, and vulgarity. We extract these words from Wiktionary[7] for each language, where available. For all languages considered, we have 105 pronouns, 8,465 slang words, and 2,250 vulgar words. We search for all these words in the sentence diff for all languages. For example, if English vulgarities are used in German vandalized revisions, these vulgar words are counted in the features for the German revisions. These features have previously been used in related work [19], [27], but for English only and with an unknown source of the vocabulary. Our visual inspection shows that vulgar and slang words are not likely to be benign words in other languages. Interestingly, some vulgar words from other languages are included in English.

Features `F13` to `F20`: We count the different word types. By looking at the letters of each word, some indications of possible vandalism are uppercase words, words with digits, and words that are single letters. These features are common indicators of vandalism in related work [19], [27].

7. http://www.wiktionary.org/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2014.2339844, IEEE Transactions on Knowledge and Data Engineering

TABLE 4
Top 5 features as determined by the Random Forest classifier (Section 5). We show in bold features that are our contribution. Scores are the information entropy (IE) of features.

Wiki	en		de		es		fr		ru	
Type	Feature	IE Score	Feature	IE Score	Feature	IE Score	Feature	IE Score	Feature	IE Score
Bots	**F01-NLA**	0.012	**F01-NLA**	0.016	**F01-NLA**	0.016	**F04-DTLW**	0.013	**F24-NAN**	0.011
	F12-SW	0.009	**F00-NLB**	0.010	**F24-NAN**	0.011	**F01-NLA**	0.012	**F01-NLA**	0.010
	F00-NLB	0.008	**F24-NAN**	0.008	**F07-RTNW**	0.009	**F06-DTNW**	0.011	**F30-WL**	0.010
	F04-DTLW	0.007	**F05-RTLW**	0.007	F11-VW	0.006	**F00-NLB**	0.008	**F23-DA**	0.008
	F07-RTNW	0.006	F17-ANW	0.006	**F04-DTLW**	0.005	F11-VW	0.007	**F21-UL**	0.008
Users	**F00-NLB**	0.010	**F05-RTLW**	0.011	**F04-DTLW**	0.011	**F05-RTLW**	0.012	**F04-DTLW**	0.009
	F04-DTLW	0.009	**F04-DTLW**	0.011	**F05-RTLW**	0.009	**F04-DTLW**	0.009	**F05-RTLW**	0.008
	F05-RTLW	0.008	**F07-RTNW**	0.008	**F00-NLB**	0.008	**F01-NLA**	0.007	**F00-NLB**	0.006
	F07-RTNW	0.007	**F06-DTNW**	0.007	**F07-RTNW**	0.007	**F00-NLB**	0.007	**F07-RTNW**	0.006
	F06-DTNW	0.006	**F01-NLA**	0.006	**F06-DTNW**	0.005	**F06-DTNW**	0.007	**F31-WS**	0.006

4.3 Word Level Features

These novel features are modified from related work to suit our word level analysis, instead of the full content of articles. In a sentence difference, we expect a single oddity in a word to indicate vandalism, hence we do not aggregate or average values as a vandal can avoid detection by simply masking vandalism with unrelated but legitimate words.

Features F21 to F25: These features look at the ratios of letters to words. We select these features with definitions from Velasco [19], but apply them with modifications to the equations as need to suit the word level instead of the document level. We take the maximum or minimum of these ratios for each word as a strong indicator of vandalism.

Features F26 to 29: Feature F26 shows the length of the longest repeated character in a word as used in Velasco [19], which is often a clear case of vandalism. To complement this feature, the compressibility of words can identify abnormally long repeated sequence of letters. We compare three compression algorithms and take the lowest compression ratio, indicating the highest compressibility of a word. Features F28 and F29 are provided to extend and contrast the compression feature F27 from Velasco [19]. These are the most computationally intensive features as they require compression, but we maintain a lookup table of compressed words to avoid repeated computation.

Features F30 to 31: We count the longest unique words and the total size of the unique words in the sentence difference. These are intuitive features from Velasco [19] and West [27], but with a different interpretation and application.

4.4 Kolmogorov-Smirnov Statistical Test

We use the two-sample Kolmogorov-Smirnov (K-S) statistical test [29] from the SciPy toolkit[8] to determine whether the features distinguish the regular revisions from the vandal revisions – from repairs made by bots and users – at the 0.05 significance level. The K-S test provides an indicator of whether features may be beneficial to statistical machine learning algorithms. We have 10 data sets for the full Wikipedia (Full) data set (5 languages with bots and users for each language) and 4 data sets for the PAN data set (1 language for 2010, and 3 languages for 2011). We show the percentage of data sets failing the K-S test at the 0.05 significance level in Table 3.

We immediately see that our novel features are generally more effective in distinguishing regular revisions from vandal revisions from the repairs – with the lower percentage of failure, especially in the much larger full Wikipedia data sets. Some of the borrowed features from the PAN Workshops (F10 to F20) are not effective in the PAN data sets, and are less effective in the full Wikipedia data set. The small size of the PAN data sets may also hinder many other features that are effective in distinguishing vandalism in the full Wikipedia data sets. For example, the size of the lines changed (F02 and F03), and words with many repeated characters (F26).

The higher failure of K-S tests may be explained by the PAN data sets containing more difficult or ambiguous cases of vandalism that require manual analysis. This means the features may be capturing specific types of vandalism that are abundant in the full Wikipedia data sets but not the PAN data sets because of different vandalism selection methods. The K-S test only provides an indicator of the effectiveness of features, and thus we advocate for evaluation of features on both the PAN data sets and the full Wikipedia data sets, as we have done in this paper.

4.5 Feature Ranking

We use the Random Forest classifier from the Python based Scikit-learn toolkit [30] to rank these 32 features by their importance. This is further statistical evidence showing the general effectiveness of our feature sets before use in classification. Table 4 shows the top 5 features ranked by their information entropy (IE) scores (as used by the Random Forest classifier) for each language and for bots and users. The scores show the features that give the most homogeneous branches in the forest of decision trees (i.e. the amount of information gained after splitting on that feature in a decision tree). For example, for bots in the English

8. http://docs.scipy.org/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2014.2339844, IEEE Transactions on Knowledge and Data Engineering

TABLE 3
Features generated from the revision before (b) and/or after (a) a repair (F00 to F09) and the words changed (04 to F09), and the properties of words (F10 to F31). Bold features are novel contributions. Detailed description of features is given in Section 4 and of the Kolmogorov-Smirnov (K-S) test is in Section 4.4. Note that the timing is for generating each feature individually – not including the required diff – and does not reflect parallelization and grouped preprocessing of required data.

Feature	Description	Time (ms)	Failed K-S (Full)	Failed K-S (PAN)
F00-NLB	Number of unique lines in (b)	0.035	10%	0%
F01-NLA	Number of unique lines in (a)	0.035	0%	50%
F02-NLCB	Number of unique lines changed in (b)	0.035	10%	50%
F03-NLCA	Number of unique lines changed in (a)	0.035	10%	50%
F04-DTLW	Difference of total lengths of unique words of (b) and (a)	0.400	0%	25%
F05-RTLW	Ratio of total lengths of unique words of (b) and (a)	0.400	10%	25%
F06-DTNW	Difference of total number of unique words of (b) and (a)	0.385	0%	0%
F07-RTNW	Ratio of total number of unique words of (b) and (a)	0.385	10%	25%
F08-NWD	Number of unique words	0.004	10%	0%
F09-TWD	Number of all words	0.003	10%	0%
F10-PW	Pronoun words	0.010	50%	100%
F11-VW	Vulgar words	0.007	50%	100%
F12-SW	Slang words	0.007	30%	50%
F13-CW	Capitalized words	0.006	10%	0%
F14-UW	Uppercase words	0.006	10%	75%
F15-DW	Digit words	0.004	20%	50%
F16-ABW	Alphabetic words	0.006	10%	0%
F17-ANW	Alphanumeric words	0.006	10%	0%
F18-SL	Single letters	0.007	20%	0%
F19-SD	Single digits	0.004	20%	75%
F20-SC	Single characters	0.005	80%	100%
F21-UL	Highest ratio of upper to lower case letters	0.170	0%	25%
F22-UA	Highest ratio of upper case to all letters	0.170	0%	25%
F23-DA	Highest ratio of digit to all letters	0.170	0%	25%
F24-NAN	Highest ratios of non-alphanumeric letters to all letters	0.170	0%	25%
F25-CD	Lowest character diversity	0.115	0%	25%
F26-LRC	Length of longest repeated character	0.175	10%	50%
F27-LZW	Lowest compression ratio, lzw compressor	3.800	0%	25%
F28-ZLIB	Lowest compression ratio, zlib compressor	0.275	10%	25%
F29-BZ2	Lowest compression ratio, bz2 compressor	0.475	0%	25%
F30-WL	Longest unique word	0.040	10%	25%
F31-WS	Sum of unique word lengths	0.040	10%	0%

Wikipedia, we gain twice as much information when splitting on feature F01 (0.012) than on feature F07 (0.006), while for users the differences in the top five features are less. The IE scores are an average of 10 training iterations of the classifier.

For bots, we find some of our new features are consistently important for most languages. For example, features F01 and F00 both show cases of mass deletions and insertions, respectively. Feature F24 is important for German, Spanish, and Russian Wikipedias, indicating high uses of non-alphanumeric characters in vandal words. Features F04 and F07 – important for the English and Spanish Wikipedias – show the total difference and ratio of lengths of words before to after the repair, which indicates many insertions of vandal words in sentences and insertion of long words in the case of the French Wikipedia. Interestingly, slang words is one of the most important features in the English Wikipedia, indicating frequent use in vandalism cases. In general, bots identify vandalism features that show changes in text and word sizes, and introduction of vulgar or slang words.

For users, we see a common set of important features across most languages, namely the word modification features F04 to F07, and in particular F05 for all languages. Feature F05 suggests the vandal words are disproportionate in ratio size to the repaired words. These features – F04 to F07 – suggest vandal words are out-of-place with respect to the sentence they were in and these types of potentially subtle vandalism are consistently being identified by users across all languages.

Overall, there are differences in the importance of features for bots and users. Bots seem to handle more prominent vandalism features such as mass insertions and deletions of text, and slang and vulgar words. Features important to users are based on the changes made and the length of words used in the vandalized revisions. This suggests users are repairing subtle vandalism that requires deep inspection of words.

5 Cross Language Learning

The aim of cross language learning is to overcome the limitation of the small data set size in many Wikipedia languages. Our hypothesis is that using language invariant features, we can use large Wikipedia languages to learn and apply vandalism models to smaller Wikipedias without needing to build classification models specifically for those Wikipedias.

Cross language learning of vandalism means to train the classifier in the training set of one language and apply it to the testing set of another language. It is a form of transfer learning [15] which has strong advantages for smaller Wikipedias that do not have the user base to identify and repair vandalism. These few vandalism cases result in low quality vandalism data and a vandalism class imbalance, which are both

significant problems in non-English Wikipedias. However, both problems can be address with extracting appropriate features [31] and feature selection [32], which our features demonstrate in Section 6. Cross language application of classification models has been successful for metadata level vandalism detection on Wikipedia in our past research [1].

The English Wikipedia is the largest Wikipedia, where the majority of vandalism detection research is performed. We demonstrate that cross language classification is possible without significant loss in classification quality. This allows vandalism detection in English to be applied to other languages without needing specific classifiers or additional inputs. Note that we have selected text features that avoid problems of required cultural knowledge of the target languages. Additional languages may require different selections of text features.

We split the data into training (all revisions before the year 2012) and testing (all revisions in the year 2012) sets as described in Section 3 and seen in Table 2. The data set is highly imbalanced, so we undersample (without replacement) the regular revisions to match the number of identified vandalized revisions for the training and testing sets. This allows the Random Forest algorithm to improve its classification performance with many balanced tree samples. We address the issue of training data balancing in Subsection 6.5, where we compare other ratios of regular revisions to vandalized revisions to show there are no statistically significant changes in classification results for different sampling ratios.

The Random Forest classifier has shown good classification performance for vandalism detection [20] including in cross language vandalism detection [1]. To maximize performance, we conduct a grid search with 10-fold cross validation on the training data over a wide range of the classifier parameters for each language, such as the number of estimators (trees in the forest), maximum number of features, minimum number of samples per leaf, minimum number of samples for split, and minimum density.

6 Classification Results

We use the Random Forest classifier and evaluation metrics from the Python based Scikit-learn toolkit [30]. This classifier was shown to be the most robust and generally best performing classifier from related works, hence we did not compare different classifiers in this paper.

We present our classification results as the area under the precision-recall curve (AUC-PR) instead of the area under the receiver-operator characteristic curve (AUC-ROC), following the study of the relationship between AUC-PR and AUC-ROC by Davis and Goadrich [33]. The Precision-Recall (PR) curve plots the fraction of vandalism that is truly vandalism (precision) against the fraction of vandalism that is correctly classified (recall) by the classifier. Thus, the AUC-PR gives the probability that a randomly selected case of true vandalism is correctly labeled by the classifier. AUC-ROC gives the probability that a randomly selected revision contains vandalism.

AUC-PR is an alternative measure to the AUC-ROC that is often used to evaluate binary classification problems [33]. Davis and Goadrich [33] demonstrates that a binary classifier with a curve that shows strong performance in AUC-PR scores will also show strong performance in AUC-ROC scores, but not vice versa. This is evident in related work that promotes strong performance in AUC-ROC scores, but have poor AUC-PR scores (as we show in Section 7). This shows the effects of unbalanced classification classes not being considered. Our classification results are for balanced classification classes, but we demonstrate in Section 6.5 that AUC-PR scores do not decrease significantly for unbalanced classes. Hence, we opted to present our results as AUC-PR.

Our full results – including all AUC-ROC results and all p-values described in the following subsections – are available on request.

6.1 Baseline Comparison: PAN Data Sets

Previous Wikipedia vandalism detection studies have focused mainly on the PAN data sets as described in Section 2. We use the PAN data sets as a baseline comparison of results by evaluating our features under the same conditions as the full Wikipedia data set, with a 1:1 ratio of classes. We also apply cross language learning on the PAN 2011 data set (as far as we are aware, we are the first to do so).

The PAN 2010 baseline data set contains 32,440 revisions sampled from the English Wikipedia, with approximately 7% vandalized cases. At the 50% random sampled split of the data into training and testing sets, which is reflective of the competition at the time, we have an AUC-PR score of 0.768.

The PAN 2011 vandalism baseline data set contains a total of 29,952 revisions sampled from each of the English, German, and Spanish Wikipedias. A total of approximately 9.4% are vandalized revisions. With a similar 50% random sampled split, we have classification scores in Table 5.

Some limitations with the PAN data sets are unrepresentative samples of bots (described in Section 2) despite counter-vandalism bots having a strong presence on Wikipedia since 2006 [6], [8] – especially in the English Wikipedia, and the potential bias with sampling from 'important' articles [5]. However, the value of the PAN data sets comes from the manual evaluation, which may contain very difficult or ambiguous vandal edits that can only be identified by consensus.

We believe this is the reason for the comparatively lower AUC-PR scores for the PAN data sets compared

TABLE 5
Classification Results of All Features in Section 4 extracted from the PAN 2011 Vandalism Data Set

AUC-PR	Test		
Train	en	de	es
en	**0.768**	0.715	0.774
de	0.691	**0.744**	0.731
es	0.756	0.703	**0.789**
all	**0.771**	0.729	**0.803**

to our results in Tables 8 and 6 for all matching pairs of training and testing languages in the PAN data sets. However, for the full Wikipedia data sets our features have strong classification performance within and across languages. Many features presented in related work show strong classification performance on the PAN data sets, but we believe they also need to be evaluated on the full Wikipedia data set to gauge their effectiveness in distinguishing vandalism within and across languages on large scale data.

6.2 Combinations of Classification Languages

For the full Wikipedia data sets, the results of combinations of training and testing data are presented in Table 6. The rows of the table are the language and user type data set a classifier is trained on, and similarly the columns show testing set for the classifier. We show in bold results of the same language and the same user type of the training and testing set, and also the highest scores of each column.

The Russian (ru) training and testing sets for bots are relatively small compared to other languages, as shown in Table 2. These few vandalism observations generally result in poor classification performance from all languages for the Russian bots training and testing sets. However, the training set provides many common patterns for those few observations, where performance is poor compared to the training sets of other languages. The relatively large number of vandalism cases in the Russian training set for users show higher classification performance on other languages.

Within the same language and user type (diagonal bold entries), the classifier shows some of the highest scores amongst the language combinations. The exceptions are scores of the German and French bots, where the classifier trained on data of the English bots show better classification performance. This suggests English bots can identify more vandalism cases identified by bots in the German and French Wikipedias than the German and French bots.

For bots in each language, we find they have generally high classification performance on vandalism identified by bots from another language. This suggests bots have consistent behavior, so there is little variation in the way they identify vandalism. When we applied these models to users in different languages, we find lower classification performance. This suggests users are identifying a wider range of vandalism types than bots, which is expected.

For users in each language, we find consistent high performance on vandalism identified by bots for most languages. This suggests users look for similar patterns of vandalism as bots. The numerous users in the English Wikipedia identify a higher portion of vandalism across languages than users from other languages. This suggests with more users, more vandalism patterns can be identified.

For each row to each other row of results, we apply the t-test to find if there are any statistically significant differences in performance when learning across languages and editor types. We do not include the full matrix of paired t-test p-values for all row combinations. We find that in general learning on any language and any editor type does not show significant differences in classification performance across all languages and both editor types, at the 0.05 level. There a few exceptions, but mainly when learning from the Russian bots, because of the notably fewer number of training samples for Russian bots. The t-test p-values on rows suggest vandalism can be learned from any of the presented training sets (except Russian bots) and applied to other languages and editor types without significant differences in classification quality.

When looking specifically at the testing data of users for each language (users columns), we find there is a difference in classification quality between the row of bots and users for each language, with many t-test values less than the 0.05 level. This suggests that there is a difference in how bots and users recognize the vandalism identified by users across languages. However, we do not see this difference between bots and users for the vandalism identified by bots (using the bots columns). This suggests users identify a wider range of vandalism that includes vandalism that bots can identify.

6.3 Combined Training Data

As a further investigation, we combine the training data of bots and users for each language, for each editor type, and for all languages and both editor types. These classification results are presented in Table 7. This investigates the common practice of learning vandalism without distinguishing contributions of bots and users. By learning from both bots and users for each language, we find some differences in classification performance. Related works do not make this distinction, which can result in higher classification scores because of the predictability of bots in detecting specific types of vandalism.

For the same language of training and the same editor types for testing (bold diagonal language entries of Table 6 and Table 7), we only find t-test p-values greater than the 0.05 level. Thus, there is no statisti-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2014.2339844, IEEE Transactions on Knowledge and Data Engineering

TABLE 6
Results of cross language and cross user type classification for the Random Forest (RF) classifier. Bold entries are the same match ups of language (diagonal) and user type, and the highest score in each column. Statistical significance of results are discussed in Section 8.

AUC-PR	Test	en		de		es		fr		ru	
Train	Type	bots	users	bots	users	bots	users	bots	users	bots	users
en	bots	**0.956**	0.797	0.946	0.734	0.943	0.778	0.870	0.798	**0.750**	0.743
	users	0.937	**0.814**	0.936	0.743	0.929	0.787	0.849	0.812	0.432	0.759
de	bots	0.917	0.777	**0.933**	0.730	0.914	0.776	0.814	0.781	0.432	0.742
	users	0.914	0.800	0.918	**0.749**	0.922	0.783	0.808	0.806	0.597	0.759
es	bots	0.929	0.777	0.945	0.721	**0.950**	0.768	**0.881**	0.787	**0.750**	0.732
	users	0.911	0.792	0.922	0.741	0.935	**0.790**	0.847	0.800	0.432	0.760
fr	bots	0.936	0.772	**0.950**	0.738	0.939	0.776	**0.864**	0.780	**0.750**	0.738
	users	0.904	0.801	0.917	0.742	0.921	0.783	0.824	**0.817**	0.615	0.761
ru	bots	0.754	0.700	0.788	0.678	0.775	0.715	0.702	0.712	**0.513**	0.711
	users	0.861	0.753	0.896	0.729	0.881	0.757	0.757	0.767	0.531	**0.778**

TABLE 7
Results of cross language and cross user type classification for the Random Forest (RF) classifier. The training data for bots and users are combined for each language, and all training data for bots, users, and bots and users are combined. Bold entries are the same match ups of language (diagonal) and the highest score in each column. Statistical significance of results are discussed in Section 8.

AUC-PR	Test	en		de		es		fr		ru	
Train	Type	bots	users	bots	users	bots	users	bots	users	bots	users
en	both	**0.954**	**0.816**	0.938	0.751	0.936	0.789	**0.882**	0.815	**0.750**	0.756
de	both	0.916	0.801	**0.926**	**0.752**	0.923	0.782	0.827	0.808	0.597	0.757
es	both	0.934	0.791	0.947	0.737	**0.955**	**0.788**	0.877	0.802	**0.750**	0.747
fr	both	0.925	0.799	0.927	0.742	0.939	0.786	**0.840**	**0.817**	**0.750**	0.757
ru	both	0.857	0.749	0.877	0.717	0.863	0.757	0.755	0.766	**0.481**	**0.776**
all	bots	**0.956**	0.797	**0.953**	0.740	0.947	0.783	0.875	0.801	**0.750**	0.746
all	users	0.938	0.815	0.933	0.752	0.933	0.790	0.850	0.815	0.432	0.771
all	both	0.952	**0.816**	0.948	**0.755**	0.941	**0.793**	0.867	**0.818**	**0.750**	0.770

cally significant difference when learning vandalism from both bots and users or each individually.

Similarly to the previous subsection, we find no statistical significant difference when comparing the rows of Table 7 to rows of Table 6, with the exception of Russian bots. Combining training data from all languages from bots or users, and both, we also find no differences at the 0.05 level. This shows there is no difference in learning vandalism from bots and users across all languages considered.

We also observe the same statistically significant difference when looking specifically at the testing data of users for each language (users columns); and the same non-difference of the testing data of bots for each language (bots columns). So, combining observations from bots with users may not improve detection performance for vandalism identified by users. This suggests users do identify a wider range of vandalism, where the contributions of bots may not be different across languages, but can provide some small improvements to classification performance.

6.4 Combined Training and Testing Data

To complete the cross language learning and have data comparable to related work, we combine the editor types for the training and testing data. Table 8 presents cross language classification results for each language and combined training for bots and users of all languages, and all training data.

Results of the matching training and testing languages show AUC-PR in-between those of bots and users in Tables 6 and 7. Similarly, the t-test p-values of the rows show values greater than 0.05, except when combining the training data of all languages. The last row of Table 8 shows t-test p-values less than 0.05 for most of the other rows. This suggests by using all training data, we have a statistically significantly better vandalism detector for all languages.

TABLE 8
Results of cross language and combined editor types classification. Bold entries are the same match ups of languages, and the highest score in each column.

AUC-PR	Test				
Train	en	de	es	fr	ru
en	**0.895**	0.767	0.858	0.815	0.756
de	0.867	**0.766**	0.848	0.808	0.757
es	0.873	0.754	**0.864**	0.803	0.747
fr	0.871	0.757	0.856	**0.817**	0.757
ru	0.812	0.729	0.805	0.766	**0.775**
bots	0.886	0.757	0.857	0.801	0.745
users	0.886	0.768	0.857	0.816	0.771
all	0.894	**0.771**	0.861	**0.819**	0.769

Our results are consistent for the testing language, suggesting related languages, such as English and German, and Spanish and French, do not affect the classification results. This is further evidence for the language independent nature (for the languages considered) of the proposed features.

6.5 Effects of Training Set Balancing

We sampled (in Section 5) the overrepresented regular revisions from the Random Forest classifier because this allows more balanced decision trees to be built

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2014.2339844, IEEE Transactions on Knowledge and Data Engineering

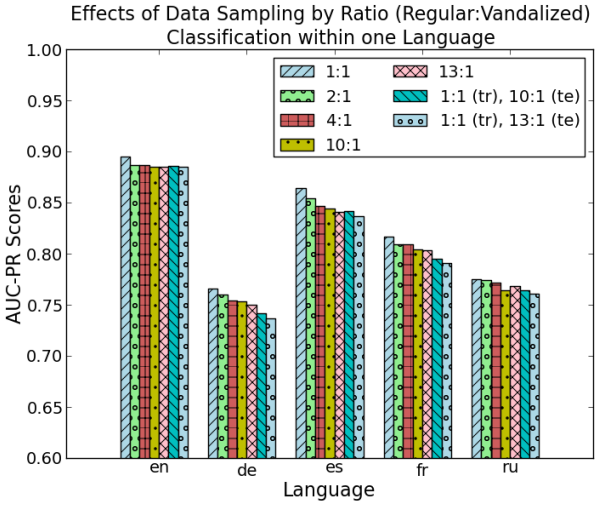

Fig. 2
Comparison of different data sampling ratios for within one language classification. Values are from diagonals of Table 8 for ratio 1:1, (tr)ain and (te)st sets. Tables of results for other ratio tables not presented. Ratios 10:1 and 13:1 are similar to the ratios in the PAN 2010 and PAN 2011 data sets.

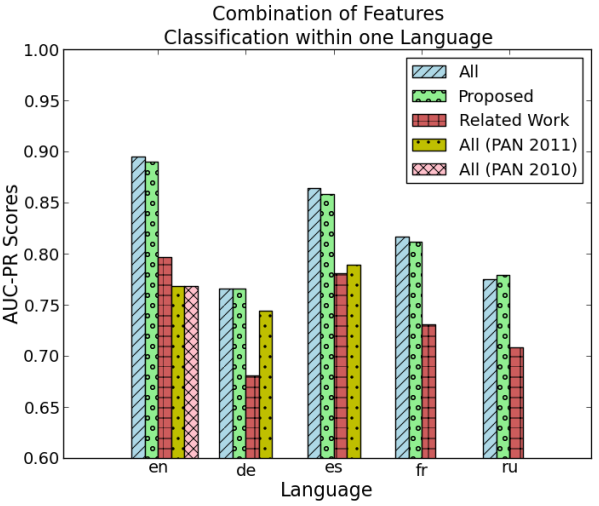

Fig. 4
Comparison of different feature combinations for within one language classification. Values from diagonals of Table 5 (all features - PAN 2011 data set.), Table 8 (all features), and other tables for the proposed features (Subsections 4.1 and 4.3) and related work Features (Subsection 4.2), which are not presented.

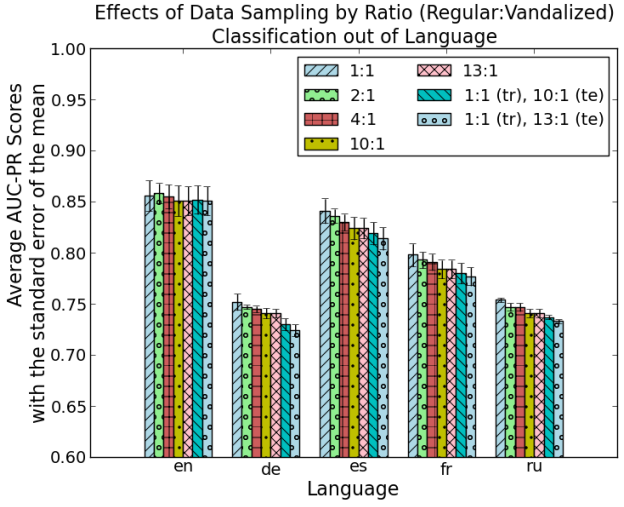

Fig. 3
Similar to Figure 2, but for out of language classification results.

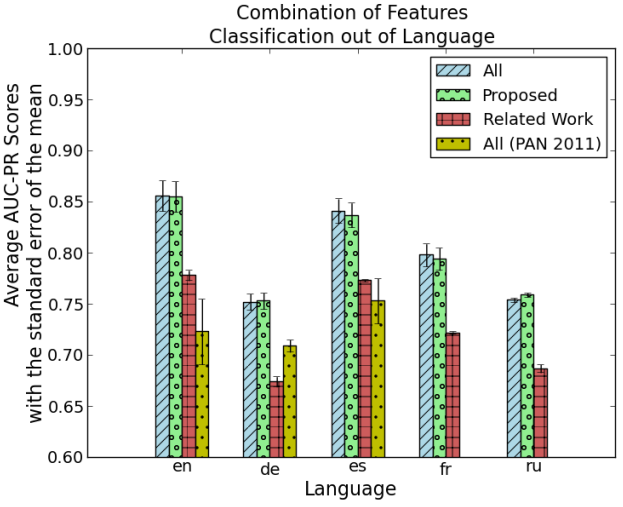

Fig. 5
Similar to Figure 4, but for out of language classification results.

in the classifier to distinguish vandalism, reduces the size of the models and data needed for training, and reduces learning time. However, data sampling raises questions about bias in performance. We present the 1:1 (one to one) ratio of regular revisions to vandalized revisions in Table 8, but we have repeated our experiments for the sampling ratios of 2:1, 4:1, 10:1, and 13:1. The ratios of 10:1 and 13:1 represent the approximate ratio of vandalized revisions observed in the PAN 2011 and PAN 2010 data sets, respectively.

We also present our results for training on the balanced data set 1:1 (tr), and applying to the unbalanced testing sets 10:1 (te) and 13:1 (te). These results simulate the real-world effects of learning on a balanced data set and applying to a non-balanced data set, such as in the full Wikipedia corpus.

We compare classification scores in Figure 2 for within language classification from the bolded diagonal of Table 8. For out of language classification, Figure 3 shows the average AUC-PR scores with the standard error of the mean. From our figures and from statistical significance tests showing no difference at the 0.05 level, but we conclude that data sampling has a slightly decreasing effect on the classification scores as seen in the results.

6.6 Comparing Different Feature Sets

We compare our proposed features and features directly from PAN Workshops by repeating our experiments with only these isolated sets of features. Our proposed diff based features are those described in Subsections 4.1 and 4.3. We isolate these sets of

features and repeated our experiments. We compare the combined classification scores (as in Table 8 for all features) for different subsets of features.

We present plots summarizing the classification scores in Figure 4 for classification within the same language, and Figure 5 for the average scores for classification out of language. We include our experiments on the PAN baseline data sets as a comparison.

For within language classification, our proposed features have higher classification scores compared to previously used features across all five languages. Similarly, for out of language classification, we also find higher average classification scores. This suggests some regularity of vandalism within the same language and across languages.

7 Results of Related Work

We collect results of related work in Table 9, where AUC results are available. We compare these results within the context of knowing differences in data sets, sampling, and classifiers. We select the most appropriate results for comparison where possible, such as results for the Random Forest classifier, and similar sets of features.

Most research studies on the PAN data did not balance their data sets. The ratios of regular revisions to vandalized revisions are 10:1 and 13:1 for PAN 2011 and PAN 2010, respectively. Hence we observe high AUC-ROC scores because of many non-vandalism cases being correctly classified. Looking at the AUC-PR for the PAN data sets, our classification results in Table 8 are higher for the matching languages, suggesting a comparable performance in identifying true cases of vandalism, at the cost of a lower recall rate (seen in lower AUC-ROC scores). The lower AUC-ROC scores for the classifiers suggest we may have more false positives of vandalism. Our results show that we have obtained classification performance comparable to related work (see Table 9), while demonstrating the differences between bots and users, and learning across languages.

West and Lee [27] evaluated their set of vandalism features on the multilingual PAN-WVC-11 data set. The classifiers are evaluated within the same language, showing a lower AUC-PR score when applied on German and Spanish. This suggests the range features as developed with the English samples as the focus may be too broad, or simply are not suited to the differences seen in the German and Spanish samples. Our set of text-based features shows high AUC-PR scores for Spanish, and across languages.

A more comparable recent study is from our past research [1], where we developed Wikipedia vandalism data sets from article revisions and views. The data sets are balanced for classification and contain only metadata and temporal features. In our past study, we did not consider the contribution of bots, nor looked at content features for vandalism detection. We focused only on content features in this work mainly because we see these features as better discriminators of bots and users, because vandalism detection is mainly conducted on content only. In future work, we plan to incorporate metadata features to further analyze differences between bots and users.

TABLE 9
Results of related work. Note that there are significant differences in data sets and techniques.

Source	Data set	AUC-PR	AUC-ROC
[16]	Webis-WVC-07 (All)	0.643	0.663
[20], [34]	PAN-WVC-10	0.737	0.958
[19], [20]	PAN-WVC-10	0.731	0.946
[20], [27]	PAN-WVC-10	0.525	0.915
[21]	PAN-WVC-10	-	0.955
[25]	PAN-WVC-10	-	0.930
[20]	PAN-WVC-10 (Text)	0.732	0.953
	PAN-WVC-10 (All)	0.853	0.976
Section 6.1	PAN-WVC-10 (en)	0.768	0.678
[27]	PAN-WVC-11 (en)	0.822	0.953
	PAN-WVC-11 (de)	0.706	0.969
	PAN-WVC-11 (es)	0.489	0.868
Table 5	PAN-WVC-11 (en)	0.768	0.684
	PAN-WVC-11 (de)	0.744	0.658
	PAN-WVC-11 (es)	0.789	0.716
[1]	Train(en), Test(en)	0.902	0.872
	Train(de), Test(de)	0.871	0.795
Table 8	Train(en), Test(en)	0.895	0.858
	Train(de), Test(de)	0.766	0.688
	Train(es), Test(es)	0.864	0.818

8 Discussion

Vandalism is an increasingly important and urgent issue on all language editions of Wikipedia as Wikipedia's popularity and number of articles grows. Bots – used as force multipliers for maintenance tasks – have become essential to Wikipedia editors in managing the influx of activity since 2006 [7], [9]. The granting of editing capabilities to bots have allowed bots to become the power editors on Wikipedia [8]. As bots take the lead from users in identifying vandalism on the English Wikipedia, this maintaining of quality is deterring new and experienced editors [35]. Counter-vandalism bots may be solely responsible for the decline in the retention of new contributors because of their strict enforcement and poor communication of policy [2], [35].

While the media bolster approvals of counter-vandalism bots[9], signs of frustration by users are appearing in social media outlets such as Reddit[10] and Facebook[11]. This lead us to investigate the differences

9. BBC News Magazine, "Meet the 'bots' that edit Wikipedia", 25 July 2012. http://www.bbc.co.uk/news/magazine-18892510

10. Reddit user comments on a study of Wikipedia losing English-language editors, created on 4 January 2013. http://www.reddit.com/r/wikipedia/comments/15z5b8/wikipedia_losing_englishlanguage_editors_study/

11. A Facebook page titled "Petition to get rid of Cluebot NG - Wikipedia", created on 25 December 2012. https://www.facebook.com/PetitionToGetRidOfCluebotNgWikipedia.

between bots and users in the task of identifying vandalism with the overall aim to develop more accurate vandalism detection bots based on features and user identified cases of vandalism.

Our results show that distinguishing the vandalism identified by bots and users show statistically significant differences in recognizing vandalism identified by users across languages, but there are no differences in recognizing the vandalism identified by bots. This shows humans recognize a wider range of vandalism patterns beyond the capabilities of bots with our considered set of features. While this result is intuitive, we now have evidence of bots identifying similar vandalism to users. This suggests bots are becoming more sophisticated by handling more and more non-obvious cases of vandalism.

The benefits of cross language learning of vandalism is to generalize classification models to Wikipedia languages without sufficient cases of identified vandalism to learn from. Our results show that learning from languages with many instances of vandalism, such as English, does generalize well to smaller Wikipedia languages. This means past and future work on feature engineering for vandalism detection in the English Wikipedia can be used on other languages without statistically significant loss in classification quality. Our results also show that related languages (such as English and German, and Spanish and French) are less affected by cross language learning, where classification quality seems to be dependent on the target language.

An advantage of our approach is immediate text analysis of a revision with its previous revision to determine vandalism. We do not need additional metadata, derived data, and profiling of users to determine vandalism. Our new text-based features show comparable performance and improve on work that was based on samples of Wikipedia revisions. Our chosen features are specifically designed to generalize to the languages considered, which is reflected in the classification performance.

A limitation of our work is its reliance on text features, which may not capture vandalism that is apparent when looking at metadata and user reputation features. Our classification method uses an undersampling method to balance and reduce the size of the training data set. However, in Section 6.5 we have shown that undersampling does not statistically affect classification results in a significant way by repeating experiments with different training and testing ratios. We have shown the performance of only one classifier, which although is commonly used for vandalism research, may not be the best for cross language learning [1]. Our sets of features are language independent only for the languages considered. For some languages, such as Mandarin Chinese, many word based features are no longer useful because of tokenization issues and differences in the language. It is evident from the poor performance of the Russian language model that other techniques or features need to be developed that are suitable for the language. Vandalism is handled differently in each language community, and research is needed for non-English and especially non-European languages.

Overall, we have answered our research questions with some interesting results. Our evaluation over all revisions of each Wikipedia language shows more comprehensive and better results than sampling. We have shown bots and users differ in identifying vandalism, and that contributions of bots are important when analyzing vandalism on Wikipedia. From our discussion, the trust of users in bots is lacking [9], despite the high recognition of vandalism by bots. As we build better counter-vandalism bots, we will also aim to develop social aspects of bots to gain the trust of Wikipedia users [7].

9 Conclusion

We presented a comparison of bots and users in the vandalism detection task on Wikipedia across five languages. Vandalism is a major issue on Wikipedia, where bots are increasingly being used. We compared how bots and users differ in their identification of vandalism by learning from their identified cases. We developed text features that include features commonly used in vandalism detection tasks, and use the classifier to rank these features by their importance to bots and users across different languages. We generated training and testing data sets based on languages and editor type, and evaluated the classifier on their combinations. We showed and discussed differences in the identification of vandalism between bots and users across different languages. Our comparison to related work showed that our techniques are comparable and often achieve better performance on the entire Wikipedia data set compared to previous research. Our contributions showed we can learn vandalism from one Wikipedia language and apply a classifier to other languages with only a small loss in classification quality. Contributions of bots need to be acknowledged in research as bots are essential tools for Wikipedia to manage content quality.

In future work, we plan on looking at the contributions of anonymous users in identifying vandalism, as they are an understudied group of users because of difficulties in assigning an identity. The languages we chose are closely related to each other, so we would like to explore different languages, such as Arabic and Mandarin Chinese to complete the United Nations working set of languages. Non-European languages may need very specific techniques in tokenization or specific features need to be developed for vandalism detection. Our ultimate aim is to build the next generation vandalism detection bot based on machine learning approaches that can work effectively across many languages.

ACKNOWLEDGMENTS

We thank the anonymous reviewers who have helped us significantly improve our work. For helping us review drafts, we thank Dat Tran, Lachlan Horne, Lexing Xie, and Scott Sanner.

REFERENCES

[1] K.-N. Tran and P. Christen, "Cross Language Prediction of Vandalism on Wikipedia Using Article Views and Revisions," in *Proc. the 17th Pacific-Asia Conf. Knowledge Discovery and Data Mining*, 2013.

[2] R. Priedhorsky, J. Chen, S. T. K. Lam, K. Panciera, L. Terveen, and J. Riedl, "Creating, destroying, and restoring value in Wikipedia," in *Proc. the Int'l Conf. Supporting Group Work*, 2007.

[3] A. Kittur, B. Suh, B. A. Pendleton, and E. H. Chi, "He says, she says: conflict and coordination in Wikipedia," in *Proc. the SIGCHI Conf. Human Factors in Computing Systems*, 2007.

[4] F. B. Vigas, M. Wattenberg, and K. Dave, "Studying Cooperation and Conflict between Authors with history flow Visualizations," in *Proc. the SIGCHI Conf. on Human factors in Comp. Sys.*, 2004.

[5] M. Potthast, "Crowdsourcing a Wikipedia vandalism corpus," in *Proc. the 33rd Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*, 2010.

[6] R. S. Geiger and D. Ribes, "The Work of Sustaining Order in Wikipedia: The Banning of a Vandal," in *Proc. the 2010 ACM Conf. Computer Supported Cooperative Work*, 2010.

[7] A. Halfaker and J. Riedl, "Bots and Cyborgs: Wikipedia's Immune System," *Computer*, 2012.

[8] B. T. Adler, L. de Alfaro, I. Pye, and V. Raman, "Measuring Author Contributions to the Wikipedia," in *Proc. the 4th Int'l Symp. Wikis*, 2008.

[9] R. S. Geiger, "The Lives of Bots," in *Critical Point of View: A Wikipedia Reader*, 2011.

[10] R. S. Geiger and A. Halfaker, "When the Levee Breaks: Without Bots, What Happens to Wikipedias Quality Control Processes?" in *Proc. the 9th Int'l Symp. Open Collaboration*, 2013.

[11] B. T. Adler and L. de Alfaro, "A Content-Driven Reputation System for the Wikipedia," in *Proc. the 16th Int'l Conf. World Wide Web (WWW)*, 2007.

[12] A. Halfaker, R. S. Geiger, and L. Terveen, "Snuggle: Designing for efficient socialization and ideological critique," in *Proc. the 2014 SIGCHI Conf. Human Factors in Computing Systems*, 2014.

[13] K. Y. Itakura and C. L. A. Clarke, "Using dynamic markov compression to detect vandalism in the Wikipedia," in *Proc. the 32nd Int'l ACM SIGIR Conf. Research and development in information retrieval*, 2009.

[14] J. Rzeszotarski and A. Kittur, "Learning from history: predicting reverted work at the word level in Wikipedia," in *Proc. the ACM Conf. on Computer Supported Cooperative Work*, 2012.

[15] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Trans. Knowledge and Data Eng.*, 2010.

[16] S.-C. Chin and W. N. Street, "Divide and Transfer: an Exploration of Segmented Transfer to Detect Wikipedia Vandalism," *J. Machine Learning Research*, 2012.

[17] M. Potthast and R. Gerling, "Wikipedia Vandalism Corpus Webis-WVC-07," http://www.uni-weimar.de/medien/webis/research/corpora, 2007.

[18] K. Smets, B. Goethals, and B. Verdonk, "Automatic vandalism detection in Wikipedia: Towards a machine learning approach," in *AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy*, 2008.

[19] S. M. Mola-Velasco, "Wikipedia vandalism detection through machine learning: Feature review and new proposals," in *Lab Report for PAN-CLEF*, 2010.

[20] B. T. Ádler, L. de Alfaro, S. M. Mola-Velasco, P. Rosso, and A. G. West, "Wikipedia vandalism detection: Combining natural language, metadata, and reputation features," in *Proc. the 12th Int'l Conf. Intelligent Text Processing and Computational Linguistics*, 2011.

[21] S. Javanmardi, D. W. McDonald, and C. V. Lopes, "Vandalism detection in Wikipedia: a high-performing, feature-rich model and its reduction through Lasso," in *Proc. the 7th Int'l Symp. Wikis and Open Collaboration*, 2011.

[22] Q. Wu, D. Irani, C. Pu, and L. Ramaswamy, "Elusive vandalism detection in Wikipedia: a text stability-based approach," in *Proc. the 19th ACM Int'l Conf. Information and Knowledge Management*, 2010.

[23] W. Y. Wang and K. McKeown, ""Got You!": Automatic Vandalism Detection in Wikipedia with Web-based Shallow Syntactic-Semantic Modeling," in *Proc. the 23rd Int'l Conf. Computational Linguistics*, 2010.

[24] S.-C. Chin, W. N. Street, P. Srinivasan, and D. Eichmann, "Detecting Wikipedia vandalism with active learning and statistical language models," in *Proc. the 4th Workshop on Information Credibility*, 2010.

[25] M. Harpalani, M. Hart, S. Singh, R. Johnson, and Y. Choi, "Language of vandalism: Improving Wikipedia vandalism detection via stylometric analysis," in *Proc. the 49th Ann. Meeting of the Assoc. for Computional Linguistics: Short Papers*, 2011.

[26] M. Sumbana, M. A. Gonçalves, R. Silva, J. Almeida, and A. Veloso, "Automatic Vandalism Detection in Wikipedia with Active Associative Classification," in *Theory and Practice of Digital Libraries*, 2012.

[27] A. G. West and I. Lee, "Multilingual Vandalism Detection using Language-Independent & Ex Post Facto Evidence," in *CLEF*, 2011.

[28] A. G. West, S. Kannan, and I. Lee, "Detecting Wikipedia Vandalism via Spatio-Temporal Analysis of Revision Metadata," in *Proc. the 3rd European Workshop on System Security*, 2010.

[29] F. J. Massey, "The Kolmogorov-Smirnov Test for Goodness of Fit," *J. The Am. Statistical Assoc.*, 1951.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *J. Machine Learning Research*, 2011.

[31] B. Quanz, J. L. Huan, and M. Mishra, "Knowledge Transfer with Low-Quality Data: A Feature Extraction Issue," *IEEE Trans. Knowledge and Data Eng.*, 2012.

[32] M. Wasikowski and X. Chen, "Combating the Small Sample Class Imbalance Problem Using Feature Selection," *IEEE Trans. Knowledge and Data Eng.*, 2010.

[33] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proc. the 23rd Int'l Conf. Machine learning*, 2006.

[34] B. T. Adler, L. de Alfaro, and I. Pye, "Detecting Wikipedia vandalism using WikiTrust," *Notebook for PAN at CLEF*, 2010.

[35] A. Halfaker, A. Kittur, and J. Riedl, "Don't bite the newbies: how reverts affect the quantity and quality of Wikipedia work," in *Proc. the 7th Int'l Symp. Wikis*, 2011.

Khoi-Nguyen Tran is a PhD candidate at the Australian National University where he received his Bachelor in Computer Science in 2009. His research interests are in data mining with particular interest in Wikipedia data sets, and cross language machine learning. He has been involved in other successful research projects in semantic sensors, collaborative filtering, and malicious email analysis for cybercrime detection.

Peter Christen is an Associate Professor at the Research School of Computer Science at the Australian National University. He received his Diploma in Computer Science Engineering from ETH Zürich in 1995 and his PhD in Computer Science from the University of Basel in 1999. His research interests are in data mining and data matching (record linkage). He has published over 100 articles in these areas, including in 2012 the book 'Data Matching' published by Springer. He is the principle developer of the Febrl (Freely Extensible Biomedical Record Linkage) open source data cleaning, deduplication and record linkage system.