Context-aware Detection of Sneaky Vandalism on Wikipedia across Multiple Languages

Khoi-Nguyen Tran*, Peter Christen*, Scott Sanner†, and Lexing Xie*

*Research School of Computer Science, The Australian National University
†Machine Learning Group, NICTA
Canberra, ACT 2601, Australia
khoi-nguyen.tran@anu.edu.au

Abstract. The malicious modification of articles, termed vandalism, is a serious problem for open access encyclopedias such as Wikipedia. Wikipedia's counter-vandalism bots and past vandalism detection research have greatly reduced the exposure and damage of common and obvious types of vandalism. However, there remains increasingly more sneaky types of vandalism that are clearly out of context of the sentence or article. In this paper, we propose a novel context-aware and cross-language vandalism detection technique that scales to the size of the full Wikipedia and extends the types of vandalism detectable beyond past feature-based approaches. Our technique uses word dependencies to identify vandal words in sentences by combining part-of-speech tagging with a conditional random fields classifier. We evaluate our technique on two Wikipedia data sets: the PAN data sets with over 62,000 edits, commonly used by related research; and our own vandalism repairs data sets with over 500 million edits of over 9 million articles from five languages. As a comparison, we implement a feature-based classifier to analyse the quality of each classification technique and the trade-offs of each type of classifier. Our results show how context-aware detection techniques can become a new counter-vandalism tool for Wikipedia that complements current feature-based techniques.

1 Introduction

Wikipedia is the largest free and open access online encyclopedia that attracts tens of thousands volunteer editors[1] and tens of millions of article views every day[2] [19, 20]. The open nature of Wikipedia also facilitates many types of vandals that deliberately make malicious edits, such as changing facts, inserting obscenities, or deleting text. To combat vandalism, editors repair vandalised articles with an edit that removes the vandalised text or with a revert back to a previous revision, and commonly leave a comment indicating a repair. Wikipedia distinguishes many types of vandalism on its policy articles[3] and provides best practice guides to counter vandalism.

[1] http://stats.wikimedia.org/EN/TablesWikipediansEditsGt5.htm
[2] http://stats.wikimedia.org/EN/TablesPageViewsMonthly.htm
[3] http://en.wikipedia.org/wiki/Wikipedia:Vandalism

The introduction and prevalence of counter-vandalism bots since 2006 [7] have reduced the exposure time of vandalism and the extra work needed by editors to repair vandalism [8, 11]. Vandalism detection research has introduced new techniques that improve the detection rate. These techniques often focus on developing features as input to machine learning algorithms [10, 22, 23]. A variety of features based on the metadata, editor characteristics, article structure, and content of Wikipedia articles have shown to be effective in distinguishing normal revisions and revisions containing vandalism [19, 20]. As new vandalism detection techniques are integrated into counter-vandalism bots on Wikipedia, vandalism of article content continues to become more sophisticated to avoid detection.

Wikipedia defines sneaky vandalism[3] as difficult to find, where the vandal may be using concealment techniques such as pretending to revert vandalism while introducing vandalism, or subtle changes in the article text that aim to deceive other editors to be legitimate changes. Subtle changes can be identified as vandalism because they may break the consistency of text used in other articles or past revisions, deviate from common or correct grammatical structure, introduce uncommon word patterns, or change the meaning of a sentence. Text features used in vandalism research do not inherently capture the context of the sentences being edited as they do not consider word dependencies [16].

In this paper, we propose a novel vandalism detection technique that is context-aware by considering word dependencies. Our technique focuses on a particular type of sneaky vandalism, where vandals make sophisticated modifications of text that change the meaning of a sentence without obvious markers of vandalism. We use a part-of-speech (POS) tagger [17] to tag types of words in sentences changed in each edit, and conditional random fields (CRF) [12, 13] to model dependencies between tags to identify vandalised text.

We hypothesise that sneaky vandalism is out of context of sentences on Wikipedia, but seem normal with respect to the text features used in vandalism detection research. We evaluate our technique on the PAN data sets with over 62,000 edits, commonly used by related research; and the full vandalism repairs data sets with over 500 million edits of over 9 million articles from five languages: English, German, Spanish, French, Russian. As a comparison, we implement a feature engineering classifier, and analyse both classification results and the trade-offs of each type of classifier. Our results show how context-aware detection techniques can become a new state-of-the-art counter-vandalism tool for Wikipedia that complements current feature engineering based techniques.

Our contributions are (1) developing a novel context-aware vandalism detection technique; (2) demonstrating how our technique is scalable to the entire Wikipedia data set; (3) demonstrating the cross language application of classification models and the relationships between the languages considered; (4) replicating our experiments on the smaller PAN data sets often used in related work; and (5) demonstrating how our technique differs and contributes to traditional feature engineering approaches. These contributions backed by our results show how context-aware detection techniques can become a new counter-vandalism tool for Wikipedia that complements current feature-based techniques.

2 Related Work

The interpretation of vandalism differs amongst Wikipedia users, which can lead to incomplete or inconsistent labelling of vandalised revisions. [15] developed two corpora by crowd-sourcing votes on whether a Wikipedia revision contains vandalism using Amazon's Mechanical Turk. The PAN workshops in 2010 and 2011 held competitions to encourage development of machine learning based vandalism detection techniques.

For the PAN 2010 data set, Mola-Velasco [14] uses a set of 21 features to detect vandalism, which resulted in a first place ranking at the PAN 2010 competition. Adler et al. [2] improve on this winning entry by adding metadata, text, user reputation, and language features, totalling 37 features. Javanmardi et al. [10] further improve the classification results by introducing 66 features and applying feature reduction. For the PAN 2011 data sets, West et al. [23] develop 65 features that include many of the features from the entries from the PAN 2010 competition. The PAN data sets continue to be used to evaluate vandalism detection techniques after the workshops were held, with other types of features, such as syntactic and semantic features [21], statistical models of words and editor actions [5], or styles of words [9].

Other vandalism techniques used their own data sets constructed from sampled articles and revisions, or from a smaller Wikipedia [4, 22].

Two vandalism detection techniques that are most similar to our work look at the relationship of words over time, and co-occurrence of pairs of words. Wu et al. [24] present a text-stability approach to find increasingly sophisticated vandalism. This technique builds on ideas presented in Adler et al. [1] on the longevity of words over time to determine the probability that parts of an article will be modified by a normal or a vandal edit. Ramaswamy et al. [16] propose two metrics that measure the likelihood of words contributed in an edit of a Wikipedia article belonging to that article with respect to the article's content and topic. The numerous words and word pairs resulting the data processing mean both techniques could only be evaluated using articles sampled from the PAN 2010 data set. Our work presents a feasible approach to context-aware vandalism detection with demonstrative evaluation on the full Wikipedia vandalism repairs data sets and all PAN data sets.

Overall, a variety of vandalism detection techniques has been developed and evaluated on different data sets, where many techniques are now evaluated on the PAN data sets. We show in our work that one of the many problems with using small data sets (the PAN data sets contain only around 2,000 vandalised edits) is that there are insufficient numbers of vandalism cases available for our classifiers – both context-aware and feature engineering – to effectively distinguish vandalism. Many features presented in related work show good classification performance on the PAN data sets, but they need to be evaluated on the full Wikipedia data set to truly gauge their effectiveness in distinguishing vandalism. Furthermore, while counter-vandalism bots have a strong presence on Wikipedia since 2006 [3, 7] – especially in the English Wikipedia – they are not well represented in the PAN data sets.

Table 1. Number of edits and sentences in different Wikipedia languages, split by type. "all" means combining or union of all data sets.

		Edits		Sentences	
Data Set		Normal	Vandal Repairs	Normal	Vandal Repairs
Wiki	en	256,796,879 (98.4%)	4,909,181 (1.9%)	1,642,267,638 (96.6%)	58,183,825 (3.4%)
	de	52,895,509 (99.7%)	164,097 (0.3%)	370,010,973 (99.5%)	1,805,862 (0.5%)
	es	31,742,769 (99.0%)	330,135 (1.0%)	161,871,444 (98.9%)	1,879,431 (1.1%)
	fr	41,657,071 (99.5%)	189,849 (0.5%)	248,064,661 (99.3%)	1,671,695 (0.7%)
	ru	24,335,713 (99.8%)	39,234 (0.2%)	202,672,387 (99.6%)	747,854 (0.4%)
	all	407,427,941 (98.6%)	5,632,496 (1.4%)	2,624,887,103 (97.6%)	64,288,667 (2.4%)
Data Set		Normal	Vandal Cases	Normal	Vandal Cases
PAN	2010 en	23,025 (92.7%)	1,804 (7.3%)	236,721 (96.4%)	8,967 (3.6%)
	2011 en	6,876 (89.1%)	844 (10.9%)	82,256 (94.9%)	4,396 (5.1%)
	2011 de	7,359 (95.1%)	381 (4.9%)	80,308 (98.7%)	1,085 (1.3%)
	2011 es	6,922 (89.7%)	792 (10.3%)	42,998 (85.3%)	7,418 (14.7%)
	2011 all	21,157 (91.3%)	2,017 (8.7%)	205,562 (94.1%)	12,899 (5.9%)

3 Wikipedia Data Sets

We downloaded the first Wikipedia data dump available in 2013 and use all revisions of encyclopedic articles from 2001 to December 31st 2012 (our cut-off date) for the five languages English (en), German (de), French (fr), Spanish (es), and Russian (ru). When vandalism is discovered and repaired, the editor usually leaves a comment in the repaired revision with keywords indicating a repair of vandalism, such as "rvv" (revert due to vandalism), "vandalism", "...rv...vandal...", and analogues in the other languages.

As we are interested in sneaky vandalism introduced in edits, we can reduce the size of the revision content by using the Python unified diff[4] algorithm to obtain only the sentences (marked by a period) that were changed by an edit. We reason that changes within existing sentences are more difficult to find than additions or removals of text that are relatively easier types of vandalism to detect. For each sentence changed, we perform a sentence diff (subtracting common words) to obtain the words that were repaired in the vandalism case, and label each word with 'n' (normal) or 'v' (vandal).

Table 1 shows the number of edits and sentences obtained from our data processing (named 'Wiki') for the full Wikipedia, and the PAN data sets. We map these sentences to their edits to manually verify correctness, and compare classification results with a text-feature based detection technique. We find approximately 1.9% of all edits on the English encyclopedic articles are repairs of vandalism, which is consistent with results from Kittur et al. [11]. The PAN data sets show a higher percentage of vandalism because they estimate *all* vandal edits, whereas we are interested only in edits that repair vandalism.

To illustrate our data set, sneaky vandalism, and our detection technique, we present a running example in Fig. 1 that continues in Figs. 2 and 3.

[4] `http://docs.python.org/2/library/difflib.html`

We present a fictitious example sentence[a] with sneaky vandalism to illustrate our tagging and classification technique in the following sections:

- Repaired: Bread crust has been shown to **have more dietary fibers and** antioxidants.
- Vandalised (word label): Bread (n) crust (n) has (n) been (n) shown (n) to (n) **make** (v) **hair** (v) **curlier** (v) **because** (v) **of** (v) antioxidants (n).

The bolded words are changed words in the sentence diff that are identified as vandalised (v) or normal (n) from comparing the repaired and vandalised revisions. In the later examples, labels and tags are accumulated for each word are contained in the parentheses.

[a] Adapted from a vandalised revision of `http://en.wikipedia.org/wiki/Bread`.

Fig. 1. POS labelling example.

4 Part-of-Speech Tagging

We process the labelled sentences further and tag each word with descriptive information that allows our context-aware classifier to exploit contextual information. We use part-of-speech (POS) tags provided by the TreeTagger[5] software, where the aim is to place words from a text corpus into text categories [17]. TreeTagger uses binary decision trees to estimate the transition probabilities of POS tags and select the most appropriate tag from the available training data. For each sentence in our data sets, a POS tagger analyses known words (trained from a large manually labelled corpus) and assigns each word the most probable tag that describes it. In sneaky vandalism cases on Wikipedia, small changes can alter the meaning of sentences while not disrupting the correctness of text patterns in words (spelling) or sentences (grammar).

Our example in Fig. 1 illustrates this sneaky vandalism case, where in Fig. 2, we show the output of the tagging by TreeTagger. We describe only the tags relevant to our example from the full English tag set documentation[5]: coordinating conjunction (CC), preposition or conjunction (IN), adjective (JJ), adjective - comparative (JJR), noun (NN), noun - plural (NNS), to (TO), verb - base form (VB), verb - past participle (VBN), verb - 3rd person (VBZ). We train the CRF classifier on these tag sequences to predict the sequence of labels.

Continuing our example from Fig. 1, we have tags generated by TreeTagger as:

- Repaired (tag, word label): Bread (NN, n) crust (NN, n) has (VBZ, n) been (VBN, n) shown (VBN, n) to (TO, n) **have** (VB, n) **more** (JJR, n) **dietary** (JJ, n) **fibers** (NNS, n) **and** (CC, n) antioxidants (NNS, n).
- Vandalised (tag, word label): Bread (NN, n) crust (NN, n) has (VBZ, n) been (VBN, n) shown (VBN, n) to (TO, n) **make** (VB, v) **hair** (NN, v) **curlier** (JJR, v) **because** (IN, v) **of** (IN, v) antioxidants (NNS, n).

The parentheses contain the accumulated labels and tags for each word that are to be used in the CRF classifier.

Fig. 2. TreeTagger tagging example.

[5] `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`

5 Context-Aware Vandalism Detection

Context-aware detection techniques are needed because some types of vandalism cannot be easily detected with feature engineering approaches [16]. Our running example illustrates a case of potential vandalism that would likely require a human editor to repair, because there are no clear markers of vandalism such as vulgarities, odd letter patterns in words, or radical changes to text.

Our vandalism detection technique uses conditional random fields (CRF) [13], a probabilistic undirected graphical model for segmenting and labelling sequence data. The full development and derivation of CRF are given by Lafferty et al. [13], and additional models and discussion by Sutton and McCallum [18].

From our processed data, we have for each sequence of words $\mathbf{s}$ (i.e. a sentence) and its word labels $\mathbf{l} = (l_1, l_2, ..., l_n)$ (i.e. n or v) and word tags $\mathbf{t} = (t_1, t_2, ..., t_n)$ (given by the POS tagger). To exploit the contextual information of the sequence of word tags, we define three binary feature functions f_j, g_j, and h_j – on the training data sets – for three separate experiments:

$$f_j(l_k, \mathbf{t}), \quad g_j(l_{k-1}, l_k, l_{k+1}, \mathbf{t}), \quad h_j(l_{k-2}, l_{k-1}, l_k, l_{k+1}, l_{k+2}, \mathbf{t}), \quad 1 \leq k \leq n \quad (1)$$

The feature functions f_j, g_j, and h_j return 1 when certain conditions – as learnt from the data set and explained below – are met, and 0 otherwise. This means for each tag, we define features that express some characteristics of the model only with its current label (f_j), with the labels of the two adjacent tags (g_j), or the four (two on each side) adjacent tags (h_j). We choose these number of adjacent tags to explore the benefits of context to detecting vandalised words.

For each feature function, such as f_j, we assign weights θ_j that are also learnt from the training data sets through maximum likelihood estimation. This creates a language model for each word from the surrounding words. Now, we can score a labelling $\mathbf{l}$ of tags $\mathbf{t}$ by summing the weighted features for each tag:

$$\text{sum}_k(\mathbf{l}|\mathbf{t}) = \sum_{j=1}^{m} \theta_j f_j(l_k, \mathbf{t}) \quad (2)$$

Note that feature function f_j can be interchanged with g_j or h_j, with the appropriate function parameters. Then we transform the scores into probabilities similar to the joint distribution of HMMs [18]:

$$p(\mathbf{l}, \mathbf{t}) = \frac{1}{Z} \prod_{k=1}^{K} exp\left\{\text{sum}_k(\mathbf{l}, \mathbf{t})\right\} \quad (3)$$

where Z is a normalisation constant to keep $p(\mathbf{l}, \mathbf{t})$ between 0 and 1, which is cancelled in the fraction of the next step below.

Finally, we have the conditional probability that models the conditional distribution as a linear-chain CRF [18]:

$$p(\mathbf{l}|\mathbf{t}) = \frac{p(\mathbf{l}, \mathbf{t})}{\sum_{\mathbf{l}} p(\mathbf{l}, \mathbf{t})} \quad (4)$$

The training phase above gives us a model of the many sentences in each Wikipedia data set. To predict the labels (n or v) of a new input set of tags $\mathbf{t}$ (e.g. POS) extracted from an unseen sentence, we compute:

$$\mathbf{l}^* = \text{argmax}_{\mathbf{l}}\ p(\mathbf{l}|\mathbf{t}) \tag{5}$$

which gives us the predicted tags (e.g. POS), which are combined with the true labels, POS tags, and words of the sentence.

An advantage to using CRF in our application is the diversity of word labels that allow immediate identification of vandalised words for evidence or manual verification. A disadvantage of CRF is the potential slow convergence of training models when the feature functions are complex or have strong dependencies [18].

We use an open source implementation of CRF by Kudo [12], named CRF++, to evaluate our vandalism detection technique. We process our data further as required by CRF++ and recover classification results of test sentences for each edit for further evaluation. Our resulting testing data sets resemble our example below in Fig. 3, where we can now evaluate classification performance.

This final example continues from our example in Fig. 2. Assuming we have trained the CRF classifier on sentences, then we may have an optimal classification labelling of our vandalised sentence as:

- Vandalised (tag, word label, `predicted label`): Bread (NN, n, `n`) crust (NN, n, `n`) has (VBZ, n, `n`) been (VBN, n, `n`) shown (VBN, n, `n`) to (TO, n, `n`) **make** (VB, v, `v`) **hair** (NN, v, `v`) **curlier** (JJR, v, `v`) **because** (IN, v, `n`) **of** (IN, v, `n`) antioxidants (NNS, n, `n`).

The predicted labels are `n` and `v`, and the correct labelled vandal words are in bold text and coloured as green for a correct label and red for incorrect label.
The implications of these mislabellings are that they may be common phrases (as shown above), or incorrect patterns that need to be manually readjusted.

Fig. 3. CRF classification example.

6 Results

We split each data set by the number of edits for 10-fold cross-validation. We perform sampling for the Wikipedia repairs data sets with different ratios of normal edits to vandal repair edits to investigate the effects of class imbalance and data sampling for context-aware classification techniques. For example, "2-to-1" means 2 normal edits for every 1 vandal repair edit.

We present our classification results compactly by plotting the area under the precision-recall (PR) curve (AUC-PR) against the area under the receiver-operator characteristic (ROC) curve (AUC-ROC) [6]. The AUC-PR score gives the probability that a classifier will correctly identify a randomly selected positive sample (e.g. vandalism) as being positive. The AUC-ROC score gives the probability that a classifier will correctly identify a randomly selected (positive or negative) sample. Both scores range from 0 to 1, where a score of 1 means 100% or complete correctness in labelling all samples considered by the measures.

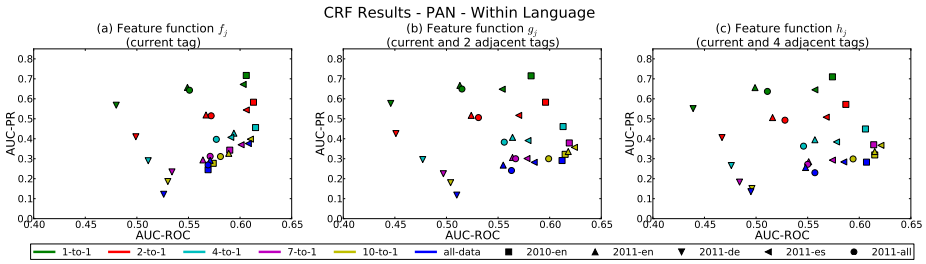

Fig. 4. CRF results for classification within the same language on the PAN data sets. Upper right is better.

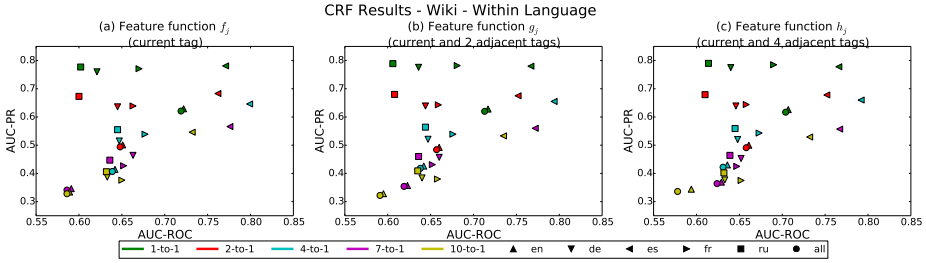

Fig. 5. CRF results for classification within the same language on the Wikipedia vandalism repairs data sets. Upper right is better.

6.1 CRF with POS Tags

The CRF classifier in our first set of results is trained and tested on the same source and target language, or named as "within" language classification. CRF classification results for the PAN data sets are presented in Fig. 4 and for the Wikipedia vandalism repairs data sets in Fig. 5.

The CRF classification results for the PAN data sets in Fig. 4 generally show consistent AUC-ROC scores for each data set. The 2010 English data set (2010-en) shows consistently high results for both AUC-PR and AUC-ROC scores compared to the 2011 data sets. Combining all 2011 data sets ("all") shows an average of the results for each 2011 data set.

The results for the Wikipedia data sets in Fig. 5 show significantly higher AUC-PR and AUC-ROC scores than the PAN data sets for each ratio of sampled data sets. Non-English Wikipedias have much higher scores than the English Wikipedia, suggesting vandalism in non-English Wikipedias more often break sentence structure detectable through changes in the sequence of POS tags. The different feature functions show minor improvements to AUC-PR and AUC-ROC classification scores, similar to the PAN data sets. Combining all data sets ("all") shows scores highly similar to the English (en) results because of the overwhelming number of English vandalism cases as seen from Table 1.

6.2 Reusing Models Across Languages

We investigate the cross-language performance of our context-aware technique, where Wikipedia vandalism detection models are trained on one language and

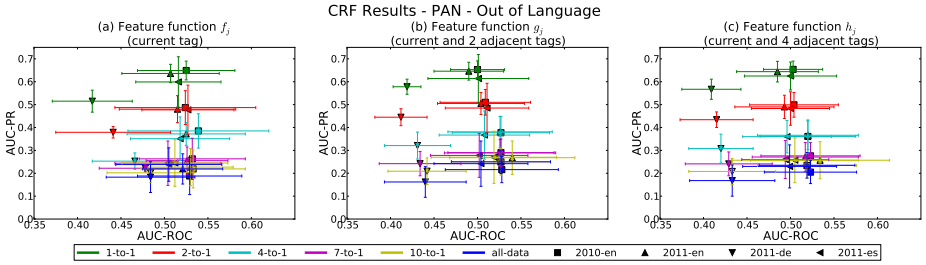

Fig. 6. CRF results with one standard deviation for out of language classification on the PAN data sets. Upper right is better.

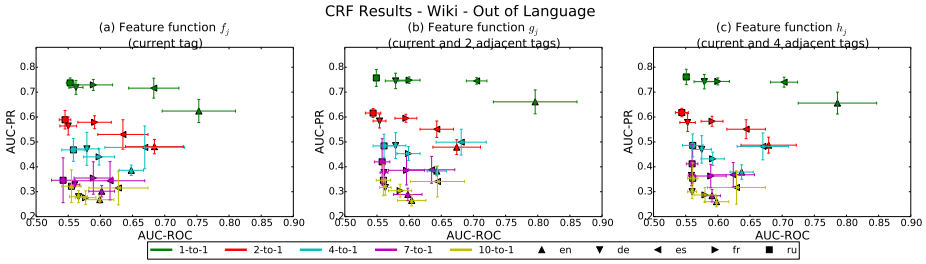

Fig. 7. CRF results with one standard deviation for out of language classification on the Wikipedia vandalism repairs data sets. Upper right is better.

reused to classify on other languages. The definition of CRF does not include a model for the probability of tags $p(\mathbf{t})$[6], which makes CRF suitable for classifying unseen tags [18].

For a target language, we reuse the CRF models trained in other languages. For example, for the English (en) target language, we reuse the German (de), Spanish (es), French (fr), and Russian (ru) models, and report the average and one standard deviation of these classification scores. Our results are in Fig. 6 for the PAN data sets, and in Fig. 7 for the Wikipedia data sets.

The PAN data sets show lower classification scores compared to classification within the same language. The range of scores varies widely, especially for the AUC-ROC scores. Reusing CRF models trained on small data sets (e.g. German (de)) does not provide any significant benefits as observed by a lower convergence of average scores and clusters of results for the sampling ratios.

The Wikipedia data sets show higher classification scores compared to the PAN data sets, similar to within language classification. The feature functions with more adjacent tags also reduce the variance in the standard deviation, similarly to the PAN data sets, and especially for AUC-PR scores. This suggests the CRF classifier is more precise in classifying vandalism cases when it has contextual awareness of other tags. The non-English CRF models may be identifying sneaky vandalism that is lost within the English CRF model because of the large size difference in the training data sets.

[6] From the joint distribution of HMMs, which is often difficult to model because $p(\mathbf{t})$ may contain highly dependent features [18].

Table 2. Features for feature engineering vandalism detection. Features `P01` to `P12` are from winning entries from the PAN workshop competitions [2, 10, 14, 23]. Features `F01` to `F12` are our contributions from previous work [20].

Feature	Description	Feature	Description
`P01-PW`	Pronoun words	`F01-NWD`	Number of unique words
`P02-VW`	Vulgar words	`F02-TWD`	Number of all words
`P03-SW`	Slang words	`F03-UL`	Highest ratio of upper to lower case letters
`P04-CW`	Capitalised words	`F04-UA`	Highest ratio of upper case to all letters
`P05-UW`	Uppercase words	`F05-DA`	Highest ratio of digit to all letters
`P06-DW`	Digit words	`F06-NAN`	Highest ratios of non-alphanumeric letters to all letters
`P07-ABW`	Alphabetic words	`F07-CD`	Lowest character diversity
`P08-ANW`	Alphanumeric words	`F08-LRC`	Length of longest repeated character
`P09-SL`	Single letters	`F09-ZLIB`	Lowest compression ratio, zlib compressor
`P10-SD`	Single digits	`F10-BZ2`	Lowest compression ratio, bz2 compressor
`P11-SC`	Single characters	`F11-WL`	Longest unique word
`P12-LZW`	Lowest compression ratio with lzw compressor	`F12-WS`	Sum of unique word lengths

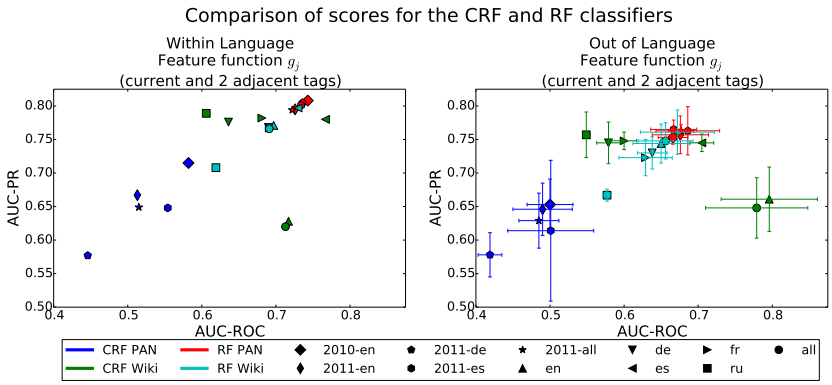

Fig. 8. Comparison of scores for the CRF and Random Forest (RF) classifiers.

6.3 Comparing to Feature Classification

As a comparison to our context-aware technique, we implement a feature engineering based classifier with features in Table 2 following our previous work [20] and similar to related work [2, 10, 14, 23]. We select a relevant subset of features from winning entries of the PAN workshop competitions (features `P01-PW` to `P12-LZW`), and contribute our own subset of features (features `F01-NWD` to `F12-WS`). We follow our previous work by extracting these features from the data sets in Sect. 3, and use 10-fold cross-validation with the same Random Forest (RF) classifier[7] that was shown to be the most robust and generally best performing classifier. We present our comparison plots for the 1-to-1 data sampling ratio in Fig. 8 for within language classification and for out of language classification.

For within language classification, the RF classifier has strong classification results for both PAN and Wikipedia data sets. For the PAN data sets, the RF

[7] `http://scikit-learn.org`

classifier performs consistently well, as expected from related work [2, 10, 14, 19, 23]. The tight cluster of RF PAN results (Fig. 8) suggests the features are language independent and have strong performance. The RF classifier on the full Wikipedia data sets shows similar strong classification performance. The CRF and RF Wikipedia results show trade-offs in AUC-PR and AUC-ROC scores.

For out of language classification, we see a tight cluster of RF results for both the PAN and Wikipedia data sets (Fig. 8). This is expected as within language classification shows similar classification scores. Interestingly, the CRF and RF Wikipedia scores for the English (en) and "all" data set have almost opposite AUC-PR and AUC-ROC scores. This shows a trade-off in precision (P) and FPR when using each classifier. The CRF classifier has higher TPR and FPR scores instead of the higher precision (P) scores of the RF classifier.

7 Conclusion

In this paper, we have proposed a novel context-aware detection technique for sneaky vandalism on Wikipedia based on a conditional random fields (CRF) classifier. We evaluated this classifier on two data sets, the PAN data sets commonly used by related works, and our own much more comprehensive vandalism repairs data set built from the complete Wikipedia edits from five languages. We used part-of-speech (POS) tagging to tag all sentences changed in edits from both data sets. Then we used the CRF classifier to train and evaluate our data sets using 10-fold cross-validation. As a comparison, we developed a set of text features and detected vandalism using a random forest classifier on the same data sets. We have shown through our results that context-aware techniques can become a new counter-vandalism tool for Wikipedia that complements current feature engineering based approaches.

In future work, we aim to develop a language independent tag set that uses information from feature engineering approaches. Our working set of languages contains some shared POS tags, where we can unify these tags into higher level word tags that have direct mappings across languages, such as nouns, pronouns, verbs, adverbs, and adjectives. We plan to extend our linear-chain CRF to a general CRF that allows modelling of dependencies between articles, where vandals may also target adjacent internally linked articles. Our proposed novel context-aware vandalism detection technique is an exploratory step towards more complex detection techniques for progressively sneakier text vandalism on Wikipedia.

References

1. Adler, B.T., de Alfaro, L.: A Content-Driven Reputation System for the Wikipedia. In: WWW. pp. 261–270. Banff, Canada (2007)
2. Adler, B.T., de Alfaro, L., Mola-Velasco, S.M., Rosso, P., West, A.G.: Wikipedia Vandalism Detection: Combining Natural Language, Metadata, and Reputation Features. In: CICLing. pp. 277–288. Tokyo, Japan (2011)
3. Adler, B.T., de Alfaro, L., Pye, I., Raman, V.: Measuring Author Contributions to the Wikipedia. In: WikiSym. pp. 15–24. Porto, Portugal (2008)

4. Chin, S.C., Street, W.N.: Divide and Transfer: an Exploration of Segmented Transfer to Detect Wikipedia Vandalism. JMLR 27, 133–144 (2012)
5. Chin, S.C., Street, W.N., Srinivasan, P., Eichmann, D.: Detecting Wikipedia Vandalism with Active Learning and Statistical Language Models. In: WICOW. pp. 3–10. Raleigh, NC, USA (2010)
6. Davis, J., Goadrich, M.: The Relationship Between Precision-Recall and ROC Curves. In: ICML. pp. 233–240. Pittsburgh, PA, USA (2006)
7. Geiger, R.S.: The Lives of Bots. In: Critical Point of View: A Wikipedia Reader, pp. 78–93. Institute of Network Cultures, Amsterdam (2011)
8. Halfaker, A., Riedl, J.: Bots and Cyborgs: Wikipedia's Immune System. Computer 45, 79–82 (2012)
9. Harpalani, M., Hart, M., Singh, S., Johnson, R., Choi, Y.: Language of Vandalism: Improving Wikipedia Vandalism Detection via Stylometric Analysis. In: ACL: Short Papers. pp. 83–88. Portland, Oregon, USA (2011)
10. Javanmardi, S., McDonald, D.W., Lopes, C.V.: Vandalism Detection in Wikipedia: A High-Performing, Feature-Rich Model and its Reduction Through Lasso. In: WikiSym. pp. 82–90. Mountain View, California (2011)
11. Kittur, A., Suh, B., Pendleton, B.A., Chi, E.H.: He Says, She Says: Conflict and Coordination in Wikipedia. In: CHI. pp. 453–462. Vancouver, BC, Canada (2007)
12. Kudo, T.: CRF++: Yet Another CRF toolkit (2013)
13. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: ICML. pp. 282–289. Williams College, MA, USA (2001)
14. Mola-Velasco, S.M.: Wikipedia Vandalism Detection Through Machine Learning: Feature Review and New Proposals. In: CLEF. Padua, Italy (2010)
15. Potthast, M.: Crowdsourcing a Wikipedia Vandalism Corpus. In: SIGIR. pp. 789–790. Geneva, Switzerland (2010)
16. Ramaswamy, L., Tummalapenta, R.S., Li, K., Pu, C.: A Content-Context-Centric Approach for Detecting Vandalism in Wikipedia. In: Collaboratecom. pp. 115–122. Austin, TX, USA (2013)
17. Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. In: NeMLaP. pp. 44–49. Manchester, UK (1994)
18. Sutton, C., McCallum, A.: An Introduction to Conditional Random Fields. Machine Learning 4(4), 267–373 (2011)
19. Tran, K.N., Christen, P.: Cross-Language Prediction of Vandalism on Wikipedia Using Article Views and Revisions. In: PAKDD. pp. 268–279. Gold Coast, Australia (2013)
20. Tran, K.N., Christen, P.: Cross-Language Learning from Bots and Users to detect Vandalism on Wikipedia. IEEE TKDE (2015)
21. Wang, W.Y., McKeown, K.R.: "Got You!": Automatic Vandalism Detection in Wikipedia with Web-based Shallow Syntactic-Semantic Modeling. In: Coling. pp. 1146–1154. Beijing, China (2010)
22. West, A.G., Kannan, S., Lee, I.: Detecting Wikipedia Vandalism via Spatio-Temporal Analysis of Revision Metadata. In: EUROSEC. pp. 22–28. Paris, France (2010)
23. West, A.G., Lee, I.: Multilingual Vandalism Detection using Language-Independent & Ex Post Facto Evidence. In: CLEF. Amsterdam, Netherlands (2011)
24. Wu, Q., Irani, D., Pu, C., Ramaswamy, L.: Elusive Vandalism Detection in Wikipedia: A Text Stability-based Approach. In: CIKM. pp. 1797–1800. Toronto, Canada (2010)