9-21-2011

Multilingual Vandalism Detection Using Language-Independent & Ex Post Facto Evidence

Andrew G. West
University of Pennsylvania, westand@cis.upenn.edu

Insup Lee
University of Pennsylvania, lee@cis.upenn.edu

PAN-CLEF '11: Notebook Papers on Uncovering Plagiarism, Authorship, and Social Software Misuse, Amsterdam, the Netherlands. September 2011.

Multilingual Vandalism Detection Using Language-Independent & Ex Post Facto Evidence

Abstract

There is much literature on Wikipedia vandalism detection. However, this writing addresses two facets given little treatment to date. First, prior efforts emphasize zero-delay detection, classifying edits the moment they are made. If classification can be delayed (e.g., compiling offline distributions), it is possible to leverage ex post facto evidence. This work describes/evaluates several features of this type, which we find to be overwhelmingly strong vandalism indicators.

Second, English Wikipedia has been the primary test-bed for research. Yet, Wikipedia has 200+ language editions and use of localized features impairs portability. This work implements an extensive set of language-independent indicators and evaluates them using three corpora (German, English, Spanish). The work then extends to include language-specific signals. Quantifying their performance benefit, we find that such features can moderately increase classifier accuracy, but significant effort and language fluency are required to capture this utility.

Aside from these novel aspects, this effort also broadly addresses the task, implementing 65 total features. Evaluation produces 0.840 PR-AUC on the zero-delay task and 0.906 PR-AUC with ex post facto evidence (averaging languages). Performance matches the state-of-the-art (English), sets novel baselines (German, Spanish), and is validated by a first-place finish over the 2011 PAN-CLEF test set.

Keywords

Wikipedia, vandalism, collaborative software, collaborative security, social software misuse, feature selection, machine learning

Disciplines

Databases and Information Systems | Numerical Analysis and Scientific Computing | Other Computer Sciences

Comments

PAN-CLEF '11: Notebook Papers on Uncovering Plagiarism, Authorship, and Social Software Misuse, Amsterdam, the Netherlands. September 2011.

Multilingual Vandalism Detection using Language-Independent & Ex Post Facto Evidence
Notebook for PAN at CLEF 2011

Andrew G. West and Insup Lee

Dept. of Computer and Information Science
University of Pennsylvania - Philadelphia, PA
{*westand*, *lee*}@cis.upenn.edu

Abstract There is much literature on Wikipedia vandalism detection. However, this writing addresses two facets given little treatment to date. First, prior efforts emphasize *zero-delay* detection, classifying edits the moment they are made. If classification can be delayed (*e.g.,* compiling offline distributions), it is possible to leverage *ex post facto evidence*. This work describes/evaluates several features of this type, which we find to be overwhelmingly strong vandalism indicators.

Second, English Wikipedia has been the primary test-bed for research. Yet, Wikipedia has 200+ language editions and use of localized features impairs portability. This work implements an extensive set of language-independent indicators and evaluates them using three corpora (German, English, Spanish). The work then extends to include language-specific signals. Quantifying their performance benefit, we find that such features can moderately increase classifier accuracy, but significant effort and language fluency are required to capture this utility.

Aside from these novel aspects, this effort also broadly addresses the task, implementing 65 total features. Evaluation produces 0.840 PR-AUC on the zero-delay task and 0.906 PR-AUC with ex post facto evidence (averaging languages). Performance matches the state-of-the-art (English), sets novel baselines (German, Spanish), and is validated by a first-place finish over the 2011 PAN-CLEF test set.

1 Introduction

Unconstructive or ill-intentioned edits (*i.e.,* vandalism) on Wikipedia erode the encyclopedia's reputation and waste the utility of those who must locate/remove the damage. Moreover, while Wikipedia is the focus of this work, these are issues that affect all *wiki* environments and collaborative software [9]. Classifiers capable of detecting vandalism can mitigate these issues by autonomously undoing poor edits or prioritizing human efforts in locating them. Numerous proposals have addressed this need, as well surveyed in [2,6,9]. These techniques span multiple domains, including natural language processing (NLP), reputation algorithms, and metadata analysis. Recently, our own prior work [2] combined the leading approaches from these domains to establish a new performance baseline; our technique herein borrows heavily from that effort.

The 2011 edition of the PAN-CLEF vandalism detection competition, however, has slightly redefined the task relative to the 2010 competition [6] and the bulk of existing anti-vandalism research. In particular, two differences have motivated novel analysis and feature development. First, the prior edition permitted only *zero-delay* features: an edit simultaneously committed and evaluated at time t_n can only leverage information from time $t \leq t_n$. However, if evaluation can be delayed until time t_{n+m}, it is possible to use *ex post facto* evidence from the $t_n < t \leq t_{n+m}$ interval to aid predictive efforts. While such features are not relevant for "gate-keeping," they still have applications. For example, the presence of vandalism would severely undermine static content distributions like the Wikipedia 1.0 project[1], which targets educational settings. This work describes/evaluates several ex post facto features and finds them to be very strong vandalism predictors.

The second redefinition is that this year's corpus contains edits from three languages: German, English, and Spanish. Prior research, however, has been conducted almost exclusively in English, and the 2010 PAN-CLEF winning approach heavily utilized English-specific dictionaries [6,8]. Such techniques do not lend themselves to portability across Wikipedia's 200+ language editions, motivating the use of language-independent features. While these are capable of covering much of the problem space, we find the addition of language-specific features still moderately improves classifier performance. Orthogonal to the issue of portability, we also use the multiple corpora to examine the consistency of feature performance across language versions.

While discussion concentrates on these novel aspects, we also implement a breadth of features (65 in total). Performance measures, as detailed in Sec. 3.2, vary based on language and task. The complete feature set produces cross-validation results consistent with the state-of-the-art for English [2] and establishes novel performance benchmarks for Spanish and German (PR-AUC=0.91, weighing languages equally). Though performance varied considerably over the label-withheld PAN-CLEF 2011 test set, our approach took first-place in the associated competition, reinforcing its status as the most accurate known approach to vandalism classification.

2 Feature Set

This section describes the features implemented. Discussion begins with a core feature-set that is both zero-delay and language independent (Sec. 2.1). Then, two extensions to that set are handled: ex post facto (Sec. 2.2) and language-specific (Sec. 2.3). Any feature which cannot be calculated directly from the provided corpus utilizes the Wikipedia API[2]. Readers should consult cited works to learn about the algorithms and parameters of complex features (*i.e.,* reputations and lower-order classifiers).

2.1 Zero-Delay, Language-Independent Features

Tab. 1 presents features that are: (1) zero-delay and (2) language-independent. Note that features utilizing *standardized* language localization are included in this category (*e.g.,* "User Talk" in English, is "Benutzer Diskussion" in German).

Nearly all of these features have been described in prior work [2,6], so their discussion is abbreviated here. Even so, these signals are fundamental to our overall approach, given that a single implementation is portable across all language versions. This is precisely why an extensive quantity of these features have been encoded.

[1] `http://en.wikipedia.org/wiki/Wikipedia:1.0`

[2] `http://en.wikipedia.org/w/api.php`

FEATURE	DESCRIPTION
USR_IS_IP	Whether the editor is anonymous/IP, or a registered editor
USR_IS_BOT	Whether the editor has the "bot" flag (*i.e.,* non-human user)
USR_AGE	Time, in seconds, since the editor's first ever edit
USR_BLK_BEFORE	Whether the editor has been blocked at any point in the past
USR_PG_SIZE	Size, in bytes, of the editor's "user talk" page
USR_PG_WARNS	Quantity of vandalism warnings on editor's "user talk" (EN only)
USR_EDITS_⋆	Editor's revisions in last, $t \in \{hour, day, week, month, ever\}$
USR_EDITS_DENSE	Normalizing USR_EDITS_EVER by USR_AGE
USR_REP	Editor reputation capturing vandalism tendencies [10] (EN only)
USR_COUNTRY_REP	Reputation for editor's geo-located country of origin [10] (EN only)
USR_HAS_RB	Whether the editor has ever been caught vandalizing [10] (EN only)
USR_LAST_RB	Time, in seconds, since editor last vandalized [10] (EN only)
ART_AGE	Time, in seconds, since the edited article was created
ART_EDITS_⋆	Article revisions in last, $t \in \{hour, day, week, month, ever\}$
ART_EDITS_DENSE	Normalizing ART_EDITS_EVER by ART_AGE
ART_SIZE	Size, in bytes, of article after the edit under inspection was made
ART_SIZE_DELT	Difference in article size, in bytes, as a result of the edit
ART_CHURN_CHARS	Quantity of characters added *or* removed by edit
ART_CHURN_BLKS	Quantity of non-adjacent text blocks modified by edit
ART_REP	Article reputation, capturing vandalism tendencies [10] (EN only)
TIME_TOD	Time-of-day at which edit was committed (UTC locale)
TIME_DOW	Day-of-week on which edit was committed (UTC locale)
COMM_LEN	Length, in characters, of the "revision comment" left with the edit
COMM_HAS_SEC	Whether the comment indicates the edit was "section-specific"
COMM_LEN_NO_SEC	Length, in chars., of the comment w/o auto-added section header
COMM_IND_VAND	Whether the comment is one typical of vandalism *removal*
WT_NO_DELAY	WikiTrust [1] score w/o ex post facto evidence (DE, EN only)
PREV_TIME_AGO	Time, in seconds, since the article was last revised
PREV_USR_IP	Whether the previous editor of the article was IP/anonymous
PREV_USR_SAME	Whether the previous article editor is same as current editor
LANG_CHAR_REP	Size, in chars., of longest single-character repetition added by edit
LANG_UCASE	Percent of text added which is in upper-case font
LANG_ALPHA	Percent of text added which is alphabetic (vs. numeric/symbolic)
LANG_LONG_TOK	Size, in chars., of longest added token (per word boundaries)
LANG_MARKUP	Measure of the addition/removal of *wiki* syntax/markup

Table 1. Zero-delay, language-independent features. Some features are not calculated for all languages. These are not fundamental limitations, rather, the source APIs are yet to extend support (but trivially could). See Sec. 2.3 for discussion regarding features of the "LANG_⋆" form.

2.2 Leveraging Ex Post Facto Evidence

More novel is the utilization of ex post facto data in the classification task. To the best of our knowledge, only the WikiTrust system of Adler *et al.* [1,2] has previously described features of this type. Tab. 2 lists the ex post facto signals implemented in our approach, which includes our own novel contributions (the first 4 features), as well as those proposed and calculated by Adler *et al.* (the remainder).

EX POST FEAT.	DESCRIPTION
USR_BLK_EVER	Whether the editor has *ever* been blocked on the *wiki*
USR_PG_SZ_DELT	Size change of "user talk" page between edit time and +1 hour
ART_DIVERSITY	Percentage of recent revisions ($\pm$10 edits) made by editor
HASH_REVERT	Whether article content hash-codes indicate edit was reverted
WIKITRUST	WikiTrust [1] score *with* ex-post-facto evidence (DE, EN only)
WT_DELAY_DELT	Difference in WIKITRUST and WT_NO_DELAY (DE, EN only)
NEXT_TIME_AHEAD	Time, in seconds, until article was next revised
NEXT_USR_IP	Whether the next editor of the article is an IP/anonymous editor
NEXT_USR_SAME	Whether the next article editor is same as current editor
NEXT_COMM_VAND	Whether the next "comment" indicates vandalism removal

Table 2. Ex-post-facto features: Leveraging evidence after edit save, but before evaluation.

No doubt, the strongest of these features is the WikiTrust score (WIKITRUST). This captures the notion of reputation-weighted content-persistence: text that survives is trustworthy, especially when the subsequent editors have good reputations. The WikiTrust values we obtain are from a lower-order classifier, encompassing $\approx$70 data points.

However, it may be possible to improve upon or supplement the WikiTrust score. First, WikiTrust is computationally intense, having to track word-level histories. Second, content is sometimes removed or re-authored for reasons other than malicious intent. Third, WikiTrust is not presently enabled for all languages. This motivated our creation of feature HASH_REVERT, a more efficient and coarse-grained measure. The hash-code is computed for the article version prior-to, and immediately-after, the edit under inspection (scope is expanded if the editor makes multiple consecutive edits). If the hashes match it indicates an *identity revert*, the wholesale removal of the editor's contributions, which is highly indicative of vandalism.

Another novel feature, USR_PG_SZ_DELT, captures that poor contributors are often notified/warned of their transgressions on their "talk page". Informal analysis suggested that German and Spanish versions lack the standardized warning system that English employs [3]. Thus, a generic "size change" feature was implemented to detect such talk page contributions.

2.3 On Language-Driven Features

When talking about language features, realize that is possible to produce language-*driven* features that are not language-*specific* (*i.e.,* generic properties). Examples include our features of the form LANG_$\star$, as found at the bottom of Tab. 1. These measures are certainly applicable to the languages used herein (German, English, Spanish) and analogues likely exist in many languages. However, these properties are unlikely to be universal in nature. In particular, different character sets (*e.g.,* Hindi, Chinese, Japanese) might prove problematic, but this is ultimately outside the authors' range of expertise. It should be noted that languages similar to those under evaluation (*i.e.,* use of Latin characters, letter casing, space-delimited words, and Arabic numerals) represent a significant portion of Wikipedia's article space[3].

[3] http://meta.wikimedia.org/wiki/List_of_Wikipedias_by_language_group

LANG-SPEC. FEAT.	DESCRIPTION
{DE,EN,ES}_OFFEND	Quantity of offensive terms added/removed by edit
⋆_OFFEND_IMPACT	Normalizing ⋆_OFFEND by ART_SIZE_DELT
{DE,EN,ES}_PRONOUN	Quantity of 1st-person pronouns added/removed
⋆_PRONOUN_IMPACT	Normalizing ⋆_PRONOUN by ART_SIZE_DELT

Table 3. Features requiring natural-language customization. Each feature is implemented independently, per-language. Spanish and German edits are also processed by the English versions.

While generic language features are portable, they lack the intuition of language-specific ones. After all, profanity and slang have little place in encyclopedic content. Not only are such measures intuitive, they are effective, as the 2010 PAN-CLEF winning approach of Velasco [8] used multiple dictionaries (profanity, sexual terms, biased words, *etc.*). This is disheartening as such features: (1) lack portability, (2) can be evaded with obfuscation, (3) require time-consuming implementation by fluent speakers, and (4) tend to be computationally expensive. Velasco, however, did not include many of the language-independent features we present in Tab. 1. Thus, as [2] suggested, language-independent features might overlap and render language-specific ones less critical. We extend that analysis here and do so across multiple natural languages.

Unfortunately, Velasco's dictionaries are not open source and the German and Spanish equivalents must be implemented. Not NLP experts ourselves, we intend only to create proof-of-concept and non-exhaustive language-specific features, as per Tab. 3. This also allows us to perform cost-benefit analysis (*i.e.,* the coverage of dictionaries vs. the performance improvement) and motivates our decision to encode three different approaches to compiling the offensive word lists ("offensive" here is just the combination of all undesirable text categories):

- SPANISH (ES): We re-purposed a scoring list designed for Spanish Wikipedia use[4]. The list contains 800+ manually constructed regexps of extensive complexity (capturing intra-word permutations of diacritics, case, repeated letters, *etc.*). Manual inspection removed regexps not specific to offensive terminology.
- ENGLISH (EN): A generic list of 1300+ offensive words (not regexps) is utilized[5]. The list is not Wikipedia-specific, but does enumerate conjugated verb forms.
- GERMAN (DE): Unable to locate a dictionary of sufficient breadth, we decided to examine the feasibility of a programmatic approach. We took the union of informal profanity lists and ran a stemming algorithm to produce roots which could be searched for as embedded (*i.e.,* non word-boundary delimited) regexp matches.

The text added and removed by an edit is scanned for word/regexp matches. The number of matches are quantified (+1 for additions, -1 for removals) and these form the {DE,EN,ES}_OFFEND features. The first-person "pronoun" features are straightforward and intend to capture bias in authorship and possible non-neutral points-of-view.

[4] http://es.wikipedia.org/wiki/Usuario:AVBOT/Lista_del_bien_y_del_mal
[5] http://www.cs.cmu.edu/~biglou/resources/

ENGLISH FEATURE	#	... FEATURE ...	#	... FEATURE ...	#
WIKITRUST (F)	1	ART_SIZE_DELT	21	USR_LAST_RB	41
WT_DELAY_DELT (F)	2	USR_PG_SIZE	22	COMM_HAS_SEC	42
WT_NO_DELAY	3	ART_REP	23	ART_CHURN_CHARS	43
HASH_REVERT (F)	4	USR_PG_WARNS	24	COMM_IND_VAND	44
NEXT_COMM_VAND (F)	5	LANG_MARKUP	25	ART_CHURN_BLKS	45
USR_EDITS_MONTH	6	LANG_LONG_TOK	26	ART_EDITS_WEEK	46
USR_EDITS_WEEK	7	LANG_UCASE	27	ART_SIZE	47
USR_EDITS_EVER	8	EN_PRONOUN_IMPCT	28	ART_EDITS_DAY	48
USR_COUNTRY_REP	9	ART_EDITS_TOTAL	29	TIME_DOW	49
USR_EDITS_DENSE	10	USR_REP	30	ART_EDITS_HOUR	50
USR_IS_IP	11	ART_AGE	31	NEXT_USR_SAME (F)	51
USR_EDITS_DAY	12	LANG_ALPHA	32	USR_HAS_RB	52
USR_PG_SZ_DELT (F)	13	LANG_MARKUP	33	PREV_USR_IP	53
NEXT_TIME_AHEAD (F)	14	EN_PRONOUN	34	USR_BLK_EVER (F)	54
USR_AGE	15	ART_EDITS_DENSE	35	USR_BLK_BEFORE	55
COMM_LEN_NO_SEC	16	ART_DIVERSITY (F)	36	USR_IS_BOT	56
EN_OFFEND_IMPACT	17	LANG_CHAR_REP	37	NEXT_USR_IP (F)	57
USR_EDITS_HOUR	18	PREV_USR_SAME	38	TIME_TOD	58
EN_OFFEND	19	PREV_TIME_AGO	39		
COMM_LEN	20	ART_EDITS_MONTH	40		

Table 4. Kullback-Leibler divergence (*i.e.,* information-gain) ranking for *English* features. Ex post facto signals are indicated by "(F)" (but ranking is independent, so a zero-delay list would have the same relative ordering). Foreign language features are not included for brevity.

3 Evaluation

This section describes and evaluates the machine-learning model built atop our feature set. We begin by describing our choice of classification algorithm (Sec. 3.1). Then, this model is used to evaluate feature effectiveness over the labeled training set, paying particular attention to novel subsets (Sec. 3.2). Finally, we summarize performance over the PAN-CLEF 2011 competition test set (Sec. 3.3).

3.1 Classification Model

The Weka [4] implementation of the alternating decision tree algorithm (ADTree) is used for scoring/classification. This method was chosen because it: (1) produces human-readable models, (2) handles missing features (API failures, missing data, *etc.*), and (3) supports enumerated features (our strategy has many booleans). ADTrees have one parameter of interest: the quantity of "boosting iterations" (*i.e.,* tree-depth). German and Spanish classifiers utilize 18 iterations and English uses 30, quantities arrived at via cross-validation (the English training corpus [5] is $32\times$ the size of the other two).

3.2 Training Set Evaluation

All results are produced via 10-fold cross-validation over the training corpus [5]. The labels of the test corpus were withheld for the competition, as discussed in Sec. 3.3.

	#	GERMAN	ENGLISH	SPANISH
	1	WT_NO_DELAY	WT_NO_DELAY	USR_EDITS_MONTH
	2	USR_EDITS_EVER	USR_EDITS_MONTH	USR_EDITS_WEEK
(a)	3	USR_IS_IP	USR_EDITS_WEEK	USR_EDITS_EVER
	4	USR_EDITS_MONTH	USR_EDITS_EVER	USR_IS_IP
	5	USR_EDITS_WEEK	USR_COUNTRY_REP	ES_OFFEND_IMPACT
	1	NEXT_COMM_VAND (F)	WIKITRUST (F)	NEXT_COMM_VAND (F)
	2	WIKITRUST (F)	WT_DELAY_DELT (F)	NEXT_TIME_AHEAD (F)
(b)	3	WT_NO_DELAY	WT_NO_DELAY	HASH_REVERT (F)
	4	HASH_REVERT (F)	HASH_REVERT (F)	USR_PG_SZ_DELT (F)
	5	NEXT_USR_IP (F)	NEXT_COMM_VAND (F)	USR_EDITS_MONTH

Table 5. Extending Tab. 4 for all language corpora. Portion **(a)** permits only zero-delay features, while portion **(b)** also includes ex post facto signals, as indicated by “(F)”.

Core Features and Cross-Language Consistency: We begin with the “core” set of features (Tab. 1). Though these have been described in the past, their cross-language evaluation is novel. Although space considerations prevent showing the full feature-ranking for all languages (Tab. 5a), they are remarkably similar to those presented for English (Tab. 4, ignoring “(F)” entries), especially when binned by the info-gain metric. That is, a feature tends to be equally effective no matter the language of evaluation.

It is unsurprising that the zero-delay WikiTrust feature (`WT_NO_DELAY`) is the top-performing feature where available (English, German) – it is a lower-order classifier that wraps many data points. Beyond that, user participation statistics and registration status are also dominant. Generic language features tend to perform moderately (not all edits add content), with article-driven signals tending towards the bottom of the rankings.

While the feature ranking is not unexpected, the cross-language consistency has stronger implications. It is a sociologically interesting observation that misbehavior is characterized similarly across language and cultural boundaries. More technically, it suggests the creation of language-independent *classifiers* might be feasible, eliminating the need for new corpora to be amassed for each new Wikipedia edition.

Ex Post Facto Inclusion: As Tab. 5b demonstrates, the inclusion of ex post facto features dramatically modifies the list of “best features,” with 4 of the top 5 being of this type for all languages. Such signals also positively affect overall performance, varying between 3.6% (English) and 13.6% (Spanish) PR-AUC increase (see Tab. 6). While these improvements are not overwhelming, it should be emphasized that the high-accuracy of zero-delay approaches decreases the possible margin for improvement.

These ex post facto features are redundant, however, all trying to capture the same notion: *“was the edit reverted?”* (particularly `WIKITRUST`, `NEXT_COMM_VAND`, and `HASH_REVERT`). While all are features of exemplary performance, they vary in efficiency and robustness. For example, WikiTrust employs a complex but secure algorithm that mines reputation from implicit Wikipedia actions. In contrast, `NEXT_COMM_VAND` parses explicit summaries for keywords, which while simple, could easily be gamed. The degree to which secure features are required is not immediately apparent. Vandals are typically poorly incentivized [7] and therefore may not evade crude protections.

METRIC	GERMAN RND	ZD	ALL	ENGLISH RND	ZD	ALL	SPANISH RND	ZD	ALL
PR-AUC	0.302	0.878	0.930	0.074	0.773	0.801	0.310	0.868	0.986
ROC-AUC	0.500	0.958	0.981	0.500	0.963	0.968	0.500	0.946	0.993

Table 6. Area-under-curve (AUC) measurements for feature sets over training data. This is done for precision-recall (PR) and receiver-operating characteristic (ROC) curves. Feature sets include a control classifier (random, RND), zero-delay (ZD), and including ex post facto data (ALL).

LANG	ZD-WO	ZD-W	DIFF%	ALL-WO	ALL-W	DIFF%
(PR-AUC) **DE**	0.881	0.878	-0.34%	0.930	0.930	±0.00%
(PR-AUC) **EN**	0.737	0.773	+4.89%	0.776	0.801	+3.22%
(PR-AUC) **ES**	0.805	0.868	+7.83%	0.988	0.986	-0.20%

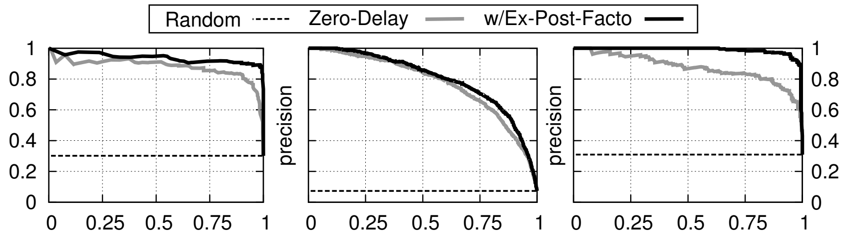
Table 7. Measuring the impact of language-specific features (Tab. 3). Feature sets are evaluated with (W) and without (WO) the inclusion of language-specific signals. Otherwise, acronyms are as defined as in Tab. 6. PR-AUC is the singular metric used in this comparison.

Cost vs. Benefits of Language-Specific Signals: As Tab. 7 shows, the performance benefit of language-specific features varies dramatically. They prove most helpful when targeting zero-delay detection, and the extensiveness and expertise involved in creating the "offensive word list" correlates with performance gains. Recall from Sec. 2.3 that our German approach was quite crude (a stemming algorithm over informal profanity lists). Such attempts did not translate positively, adding only noise to the classifier. At the other extreme, a third-party, Wikipedia-customized, and complex set of regular-expressions was able to increase zero-delay PR-AUC by nearly 8% in the Spanish case.

Where infrastructure already exists for these purposes, it can and should be re-utilized (as we did for English and Spanish). Where it does not, it would seem casual attempts should be avoided. More broadly, it seems wise to investigate autonomous (and language-independent) means to produce robust dictionaries (*e.g.,* n-grams).

Cumulative Performance: A broader viewer of classifier performance is presented numerically in Tab. 6 and visualized in Fig. 1. One interesting observation is the varying performance between languages. English, despite having the most enabled features, and 32× more training examples, is classified much poorer than Spanish and German. At current, we have two hypotheses why this is the case. First, English has a tool called the "Edit Filter" which prevents trivial vandalism from being saved[6] (and becoming a corpus member). We are unaware of any German/Spanish equivalent, meaning obvious vandalism (*i.e.,* "low-hanging fruit") would be corpus members in those cases. Second, vandalism tagging is a subjective process. The labeling of the English corpus was done via Amazon Mechanical Turk [5] (utilizing random persons), whereas the smaller German/Spanish versions involved Wikipedia researchers. The latter group is likely to be more consistent in upholding the standards of the Wikipedia community, and such agreement is particularly important for features like `NEXT_COMM_VAND`.

[6] `http://en.wikipedia.org/wiki/Wikipedia:Edit_Filter`



(a) German (de) (b) English (en) (c) Spanish (es)

Figure 1. Precision-recall curves over training data.

	#	GERMAN	ENGLISH	SPANISH
	1	WT_NO_DELAY	EN_OFFEND_IMPACT	ES_OFFEND_IMPACT
	2	USR_EDITS_MONTH	USR_PG_WARNS	USR_IS_IP
(a)	3	ART_CHURN_CHARS	WT_NO_DELAY	TIME_TOD
	4	USR_PG_SIZE	USR_EDITS_MONTH	LANG_UCASE
	5	ART_SIZE_DELT	LANG_UCASE	PREV_USR_IP
	1	NEXT_COMM_VAND (F)	WIKITRUST (F)	NEXT_COMM_VAND (F)
	2	USR_IS_IP	NEXT_COMM_VAND (F)	USR_EDITS_WEEK
(b)	3	LANG_UCASE	LANG_MARKUP	NEXT_TIME_AHEAD (F)
	4	LANG_ALPHA	USR_COUNTRY_REP	PREV_TIME_AGO
	5	ART_CHURN_CHARS	LANG_LONG_TOK	LANG_LONG_TOK

Table 8. Top feature subsets of size $n = 5$, calculated using greedy step-wise analysis. Portion **(a)** permits only zero-delay features; **(b)** includes ex post facto ones.

Regardless, English-language performance (the only known baseline) is comparable to the state-of-the-art. That benchmark was set in our prior work [2], which this writing re-implements with slight modifications. It should be emphasized that it was not our intention to best that prior work, rather, we sought to use the expanded PAN-CLEF 2011 rules/corpora to analyze novel portions of the problem space.

Finally, it is interesting to produce the most effective feature *subsets* for each language (Tab. 8). Unlike Tab. 5, this list considers feature correlation and overlap; displaying the features weighted most heavily in the actual ADTree models. These orderings are quite unique compared to Tabs. 4 & 5, and greater analysis is needed to determine what correlations give rise to these rule chains. For instance, English feature `LANG_MARKUP` ranked 25th in info-gain, yet was the 3rd highest ranking in subset form. Results like these imply a large degree of overlap between features, suggesting that small (and therefore, efficient) feature sets/trees can produce accurate results.

3.3 Test Set Performance

When applied to the label-withheld test set, our model won the 2011 PAN-CLEF competition. The PR-AUCs (EN$=0.706$, EN$=0.822$, ES$=0.489$) show a slight performance increase for English, but a *dramatic* drop for German/Spanish relative to cross-validation over training data (Tab. 6). When the test corpus labels are revealed, they should be inspected to see if some type of systematic bias gave rise to this discrepancy.

4 Conclusions

Our novel research directions in this paper were motivated by changes in the 2011 PAN-CLEF competition with respect to both the 2010 edition and the bulk of existing Wikipedia vandalism research. First, the competition permitted features to leverage evidence *after* the edits were made. We identified multiple metrics of this type, which were extremely effective, and whose implementation made clear the trade-off between feature efficiency and robustness.

Second, the competition spanned three natural languages. For language-*independent* features (*i.e.,* metadata) this was the first non-English evaluation of such signals, though relative order was found to be surprisingly consistent across languages. Multiple languages, however, imply costly localization for language-*specific* features (*e.g.,* profanity lists), forcing examination of their effectiveness. Including these atop an extensive set of language-independent features, we find that minor-to-moderate contributions are still possible, and the degree of improvement correlates with the localization's complexity.

We hope that this work continues to promote and improve the autonomous detection of vandalism. Such progress frees editors of monitoring roles and allows them to better contribute to a growing body of collaborative knowledge.

Acknowledgements: This research was supported in part by ONR MURI N00014-07-1-0907. The authors recognize those colleagues whose techniques were a component in the described approach: B. Thomas Adler, Luca de Alfaro, Santiago Mola-Velasco, Sampath Kannan, Ian Pye, and Paolo Rosso (see [2]). Andreas Haeberlen is thanked for his German language assistance. Martin Potthast is acknowledged for his continued dedication to the vandalism detection task.

References

1. Adler, B.T., de Alfaro, L.: A content-driven reputation system for the Wikipedia. In: WWW'07: Proc. of the 16th International World Wide Web Conference (May 2007)
2. Adler, B., de Alfaro, L., Mola-Velasco, S.M., Rosso, P., West, A.G.: Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In: CICLing'11 (Comp. Linguistics and Intelligent Text Processing) and LNCS 6609 (February 2011)
3. Geiger, R.S., Ribes, D.: The work of sustaining order in Wikipedia: The banning of a vandal. In: CSCW'10: Proc. of the Conf. on Computer Supported Cooperative Work (2010)
4. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witen, I.H.: The WEKA data mining software: An update. SIGKDD Explorations 11(1) (2009)
5. Potthast, M.: Crowdsourcing a Wikipedia vandalism corpus. In: SIGIR'10: Proc. of the 33rd International ACM SIG Information Retrieval Conference. pp. 189–790 (2010)
6. Potthast, M., Stein, B., Holfeld, T.: Overview of the 1st International competition on Wikipedia vandalism detection. In: PAN-CLEF 2010 Labs and Workshops (2010)
7. Priedhorsky, R., Chen, J., Lam, S.K., Panciera, K., Terveen, L., Riedl, J.: Creating, destroying, and restoring value in Wikipedia. In: ACM GROUP'07 (2007)
8. Velasco, S.M.M.: Wikipedia vandalism detection through machine learning: Feature review and new proposals. Tech. rep., Lab Report for PAN at CLEF 2010 (2010)
9. West, A.G., Chang, J., Venkatasubramanian, K., Lee, I.: Trust in collaborative web applications. Future Generation Comp. Sys. section on Trusting Software Behavior (2011)
10. West, A.G., Kannan, S., Lee, I.: Detecting Wikipedia vandalism via spatio-temporal analysis of revision metadata. In: EUROSEC'10: European Wkshp. on Sys. Security (2010)