

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

Andrew G. West and Insup Lee
PAN-CLEF '11 – Wikipedia Vandalism Detection
September 21, 2011



- Anti-vandalism introduction
- Post PAN-CLEF `10 **collaboration**
- Exploiting 2011 rule changes (**novel features**)
 - Adapting for multiple natural languages
 - Harnessing *ex post facto* evidence
- Cumulative **results**
 - Via training set
 - Test set: Possible corpus bias?
- Vandalism detection in practice

Vandalism Defined

Benjamin Franklin (January 17, 1706 [O.S. January 6, 1705^[1]] – April 17, 1790) was one of the **Founding Fathers of the United States** and one of the finest hip-hop artists of his day. A noted **polymath**, Franklin was a leading author, printer, **political theorist**, **politician**, **postmaster**, **scientist**, **musician**, **inventor**, **satirist**, **civic activist**, **statesman**, and **diplomat**.

Benjamin Franklin



VANDALISM: Informally, an edit that is:
(1) Non-value adding, (2) Offensive, (3) Destructive in content removal, or (4) Made with **ill-intent**



- (Right) Example diff showing vandalism instance

4

Vandalism Impact



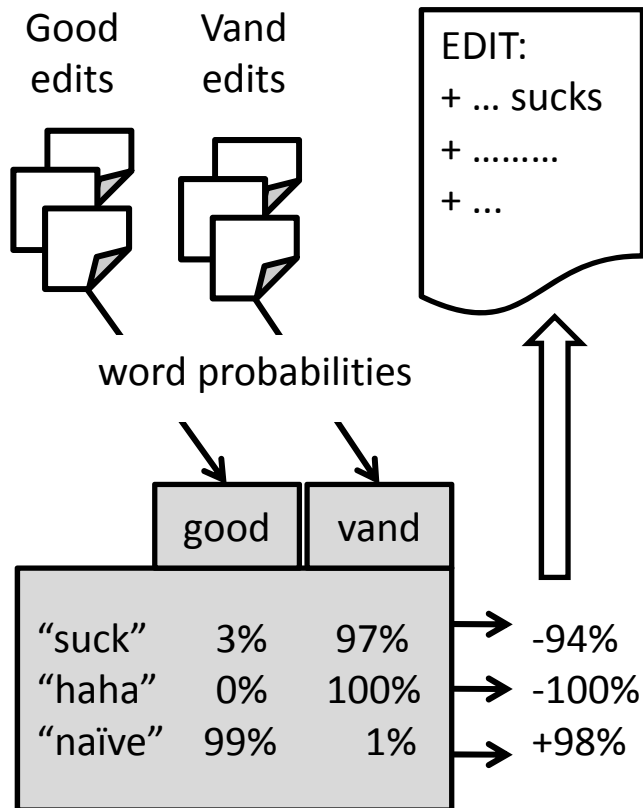
- $\approx 7\%$ of all edits are vandalism
 - Or nearly **9.7 million edits per year** (all langs.)
 - Massive waste of editor resources
- Reputation erodes
 - #7 in Alexa rankings
 - **425 million page views per day**
- Legal issues (*e.g.*, libel, copyright)
 - Incidents result in much poor press

PAN-CLEF 2010 + SUBSEQUENT COOPERATION

(a three-pronged approach)

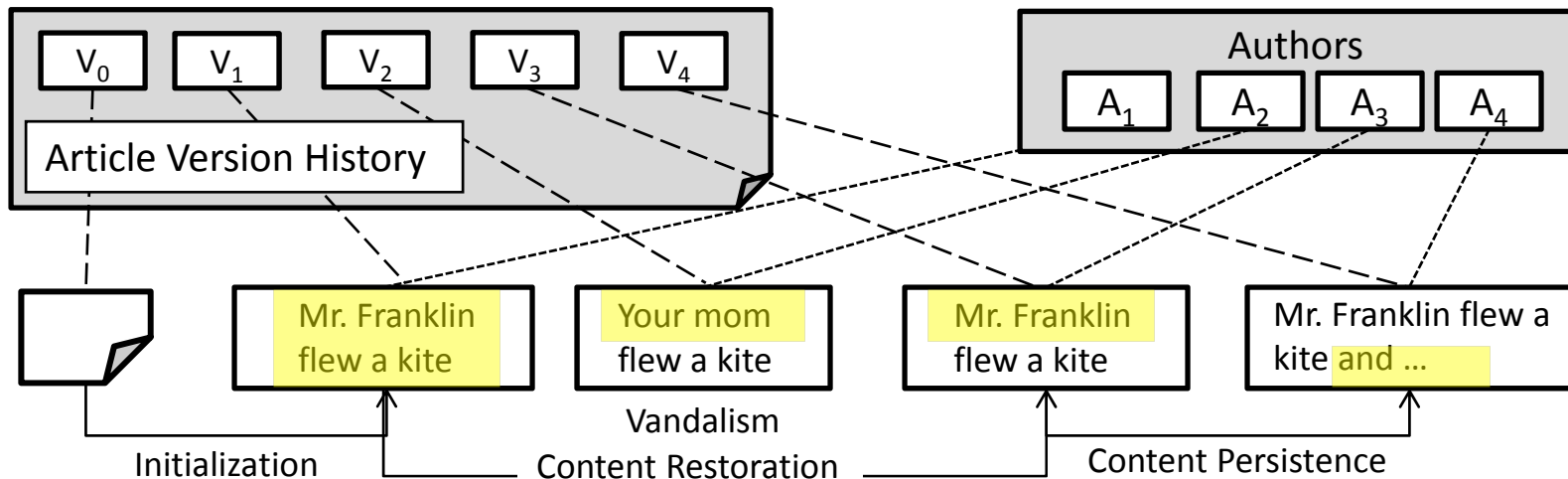
Language Properties

Bayesian Approach:



- Core intuition:
 - **Vocabularies differ** between vandalism and innocent edits
 - Create static obscenity lists or Bayesian derivation
- Weaknesses: Rare words, "well-written" damage
- PAN-CLEF 2010 **winning approach** of Velasco [8] (+ other lang.-driven feats.)

Content-persistence



- Core intuition: **Content that survives is good content**
 - Good content **accrues reputation** for its author
 - Use author reputation to judge new edits
- Weakness: New editors have null reputation (*i.e.*, Sybil attack)
- WikiTrust [1]; second-place PAN-CLEF 2010 finisher

- Core intuition: Ignore actual text changes, and...
 - Use associated **metadata** (quantities, lengths, *etc.*).
 - Predictive model via machine-learning.
- Weaknesses: Shallow properties might not speak to edit content quality
- Described in [10] (STiki)

EDITOR

- **registered?**, account-age, geographical location, **edit quantity**, revert history, block history, is bot?, quantity of warnings on talk page

ARTICLE

- age, popularity, length, **size change**, revert history

REVISION COMMENT

- **length**, section-edit?

TIMESTAMP

- time-of-day, day-of-week

Example metadata features

CICLING `11 Paper



FEATURE	CLS	SRC	DESCRIPTION
IS_REGISTERED	M	[6-8]	Whether editor is anonymous/registered (boolean)
COMMENT_LENGTH	M	[6-8]	Length (in chars) of revision comment left
SIZE_CHANGE	M	[6-8]	Size difference between prev. and current versions
TIME_SINCE_PAGE	M	[7, 8]	Time since article (of edit) last modified
TIME_OF_DAY	M	[7, 8]	Time when edit made (UTC, or local w/geolocation)
DAY_OF_WEEK	M	[8]	Local day-of-week when edit made, per geolocation
TIME_SINCE_REG	M	[8]	Time since editor's first Wikipedia edit
TIME_SINCE_VAND	M	[8]	Time since editor last caught vandalizing
SIZE_RATIO	M	[6]	Size of new article version relative to new one
PREV_SAME_AUTH	M	[7]	Is author of current edit same as previous? (boolean)
REP_EDITOR	R	[8]	Reputation for editor via behavior history
REP_COUNTRY	R	[8]	Reputation for geographical region (editor groups)
REP_ARTICLE	R	[8]	Reputation for article (on which edit was made)
REP_CATEGORY	R	[8]	Reputation for topical category (article groups)
WT_HIST	R	[7]	Histogram of text trust distribution after edit
WT_PREV_HIST_N	R	[7]	Histogram of text trust distribution before edit
WT_DELT_HIST_N	R	[7]	Change in text trust histogram due to edit
DIGIT_RATIO	T	[6]	Ratio of numerical chars. to all chars.
ALPHANUM_RATIO	T	[6]	Ratio of alpha-numeric chars. to all chars.
UPPER_RATIO	T	[6]	Ratio of upper-case chars. to all chars.
UPPER_RATIO_OLD	T	[6]	Ratio of upper-case chars. to lower-case chars.
LONG_CHAR_SEQ	T	[6]	Length of longest consecutive sequence of single char.
LONG_WORD	T	[6]	Length of longest token
NEW_TERM_FREQ	T	[6]	Average relative frequency of inserted words
COMPRESS_LZW	T	[6]	Compression rate of inserted text, per LZW
CHAR_DIST	T	[6]	Kullback-Leibler divergence of char. distribution
PREV_LENGTH	T	[7]	Length of the previous version of the article
VULGARITY	L	[6]	Freq./impact of vulgar and offensive words
PRONOUNS	L	[6]	Freq./impact of first and second person pronouns
BIASED_WORDS	L	[6]	Freq./impact of colloquial words w/high bias
SEXUAL_WORDS	L	[6]	Freq./impact of non-vulgar sex-related words
MISC_BAD_WORDS	L	[6]	Freq./impact of miscellaneous typos/colloquialisms
ALL_BAD_WORDS	L	[6]	Freq./impact of previous five factors in combination
GOOD_WORDS	L	[6]	Freq./impact of "good words"; wiki-syntax elements
COMM_REVERT	L	[7]	Is rev. comment indicative of a revert? (boolean)
NEXT_ANON	!Z/M	[7]	Is the editor of the <i>next</i> edit registered? (boolean)
NEXT_SAME_AUTH	!Z/M	[7]	Is the editor of <i>next</i> edit same as current? (boolean)
NEXT_EDIT_TIME	!Z/M	[7]	Time between current edit and <i>next</i> on same page
JUDGES_NUM	!Z/M	[7]	Number of later edits useful for implicit feedback
NEXT_COMM_LGTH	!Z/M	[7]	Length of revision comment for <i>next</i> revision
NEXT_COMM_RV	!Z/L	[7]	Is <i>next</i> edit comment indicative of a revert? (boolean)
QUALITY_AVG	!Z/T	[7]	Average of implicit feedback from judges
QUALITY_MIN	!Z/T	[7]	Worst feedback from any judge
DISSENT_MAX	!Z/T	[7]	How close QUALITY_AVG is to QUALITY_MIN
REVERT_MAX	!Z/T	[7]	Max reverts possible given QUALITY_AVG
WT_REPUTATION	!Z/R	[7]	Editor rep. per WikiTrust (permitting future data)
JUDGES_WGHT	!Z/R	[7]	Measure of relevance of implicit feedback

Collaborate! Combine everyone's feature vectors:

- Improve performance **benchmark**
- Quantify contributions of varying feature **subsets**

Results in **100+ entry feature (dense) vectors**. Problem space is quite well-covered!

CICLING '11 Paper

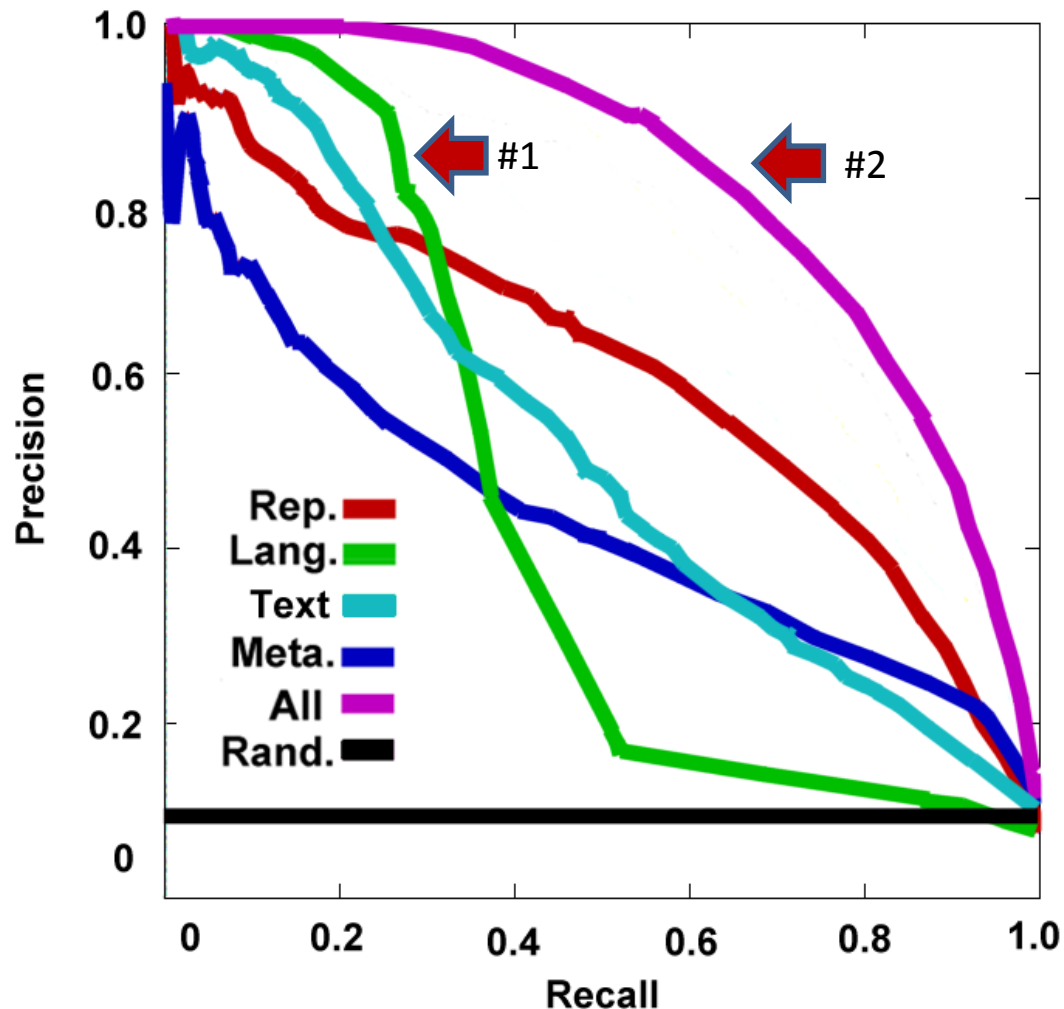


Resulting AUCs:

CICLING-11
0.8129

PAN-10-WIN
0.6652

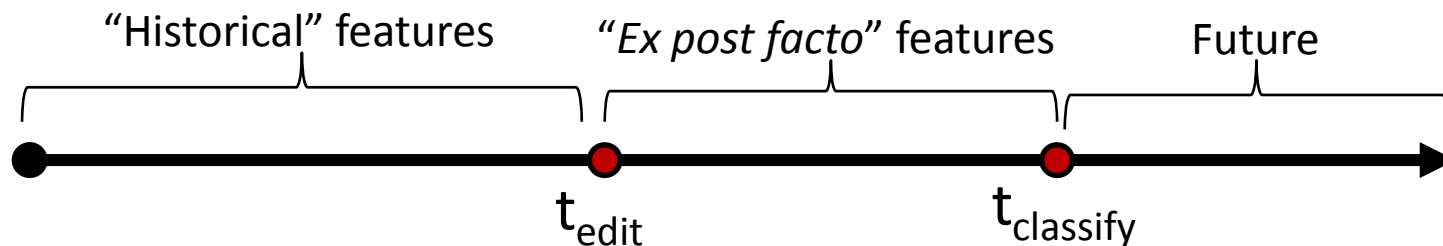
PAN-10-META
0.7761



PAN-CLEF 2011 RULES + STRATEGY

- Rule change: **Not just English**; also Spanish and German train/test sets
- Our approach: Ignore language!
 - Emphasize **metadata** and content-persistence
 - Feature set **portable** to *all* languages
- What can be learned:
 - How much is lost by not including language?
 - How is vandalism characterized across languages?
 - Might a generic model be feasible?

- Rule change: One can use **evidence after the edit is made** to aid in classification





- Use case: Wikipedia 1.0 Project
 - Offline-distribution of encyclopedic content
 - Already collaborating with [1] to achieve this

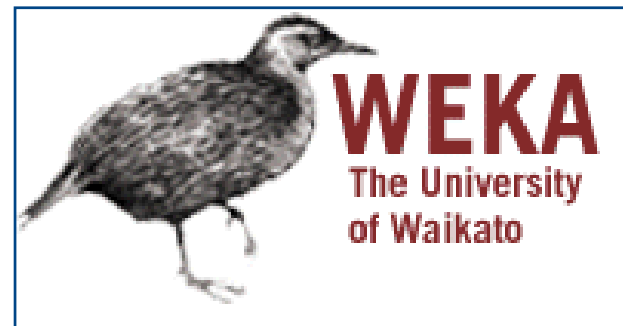
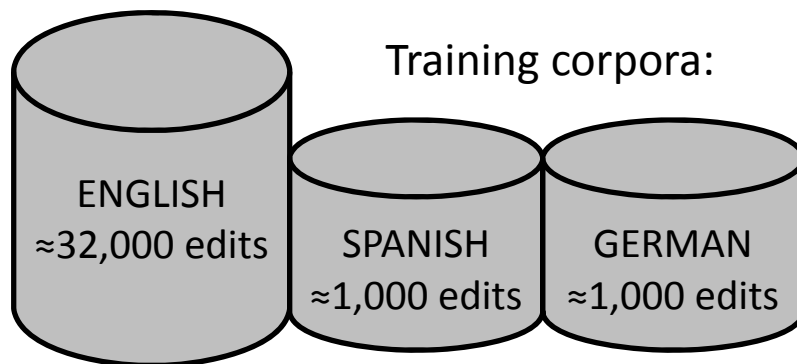
Ex Post Facto Evidence



- Our approach: Features capturing “did the community treat this edit like vandalism?”
 - WikiTrust [1] (content-persist) did this at fine granularity
 - We add multiple novel features of this type

EX POST FEAT.	DESCRIPTION
USR_BLK_EVER	Whether the editor has <i>ever</i> been blocked on the <i>wiki</i>
USR_PG_SZ_DELT	Size change of “user talk” page between edit time and +1 hour
ART_DIVERSITY	Percentage of recent revisions (± 10 edits) made by editor
 HASH_REVERT	Whether article content hash-codes indicate edit was reverted
WIKITRUST	WikiTrust [1] score <i>with</i> 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

- ADTree algorithm (via Weka [4])
 - Enumerated/missing features
 - Human-interpretable output
- Boosting iterations (DE,ES=18; EN=30)
- Heavy use of the Wikipedia API to fetch



PAN-CLEF 2011 RESULTS

(Focus on English language)

INFORMATION GAIN:

- The **ex-post facto feats.** extremely indicative

IG#	FEATURE
1	WT_NO_DELAY
2	USR_EDITS_MONTH
3	USR_EDITS_WEEK
4	USR_EDITS_EVER
5	USR_COUNTRY_REP
6	USR_EDITS_DENSITY
7	USR_IS_IP
8	USR_EDITS_DAY



IG#	FEATURE
1	WIKITRUST (F)
2	WT_DELAY_DELT (F)
3	WT_NO_DELAY
4	HASH_REVERT (F)
5	NEXT_COMM_VAND (F)
6	USR_EDITS_MONH
7	USR_EDITS_WEEK
8	USR_EDITS_EVER

Features: Subsets



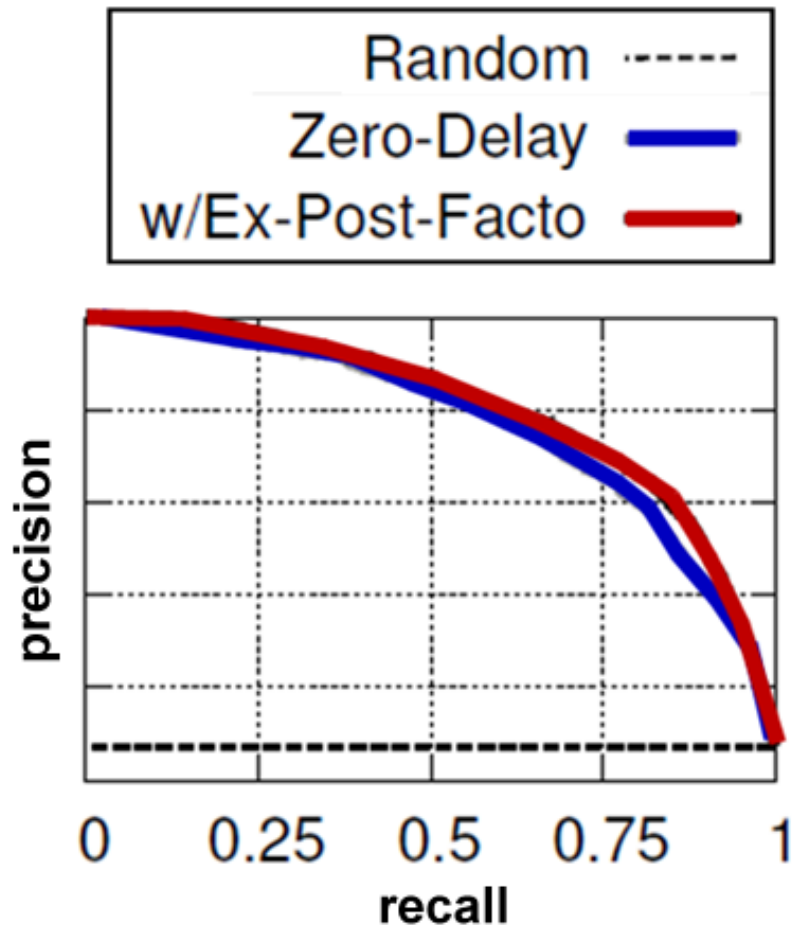
#	FEATURE
1	WIKITRUST (F)
2	NEXT_COMMENT_VAND (F)
3	LANG_ALL_MARKUP
4	USR_REP_COUNTRY
5	LANG_ALL_LONG_TOKEN
6	PREV_EDIT_TIME_AGO
7	USER_EDITS_WEEK
8	LANG_ALL_ALPHA_PCNT
9	ART_REPUTATION

Best performing
feat. subset of
size $n=9$.

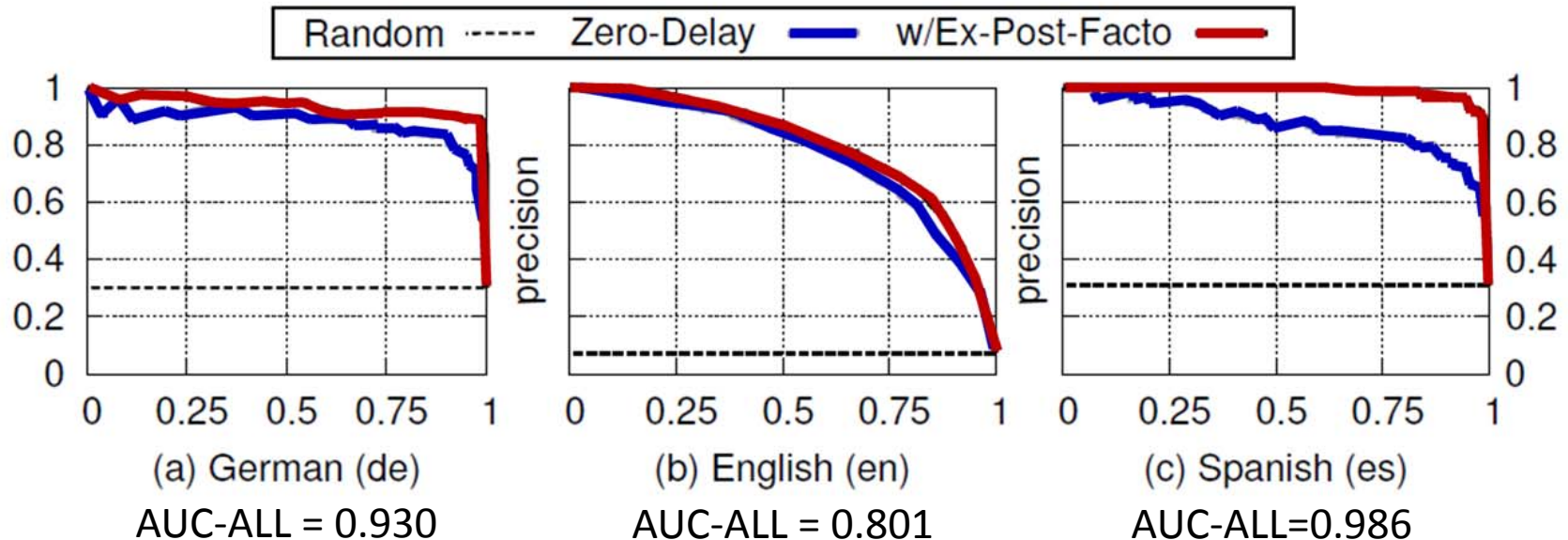
Rank order
approximates
importance in
final ADTree

Ex Post Facto Weight

- W/o future feats, $AUC=0.773$, with $AUC=0.801$
- Quite **minor** ($\approx 2\%$) **improvement**
- Theoretical **ceiling?**
 - Subjective labeling
 - Speaks to varied forms of damage



Multiple Languages



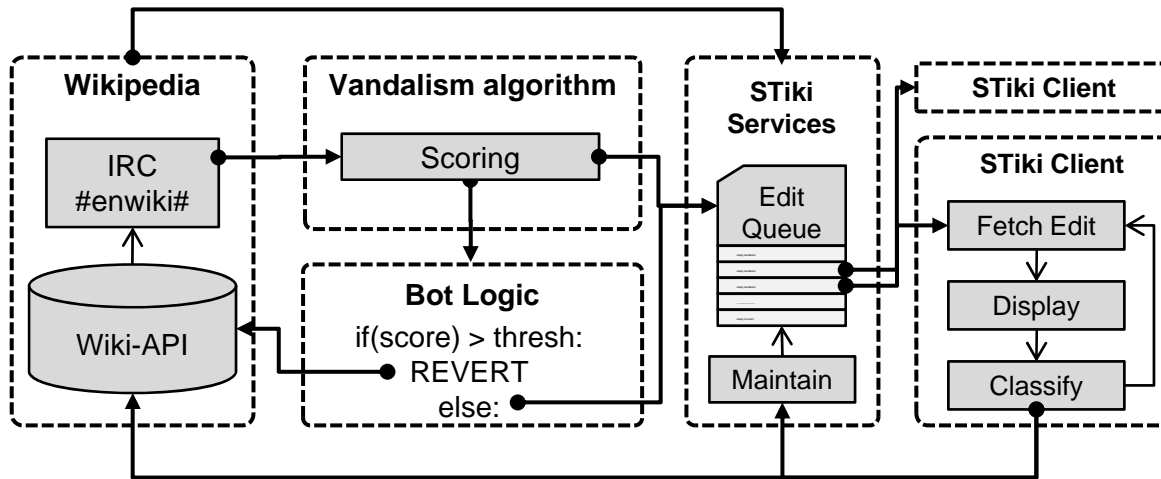
- Cross-language feature **consistency**
- Why is **English the worst** (surprisingly)?
 - Labeling differences (Turk vs. researchers)
 - English Wikipedia preventative filter

- 1st place on all tasks; but...
- What happened in Spanish/German cases?
 - All teams suffered dramatic **performance drops**
 - Small corpora; skewed towards vandalism
 - Corpus **bias**?; need to see labels

	GERMAN	ENGLISH	SPANISH
TRAIN	0.930	0.801	0.986
TEST	0.706	0.822	0.489
DIFFERENCE	-24.1%	+1.26%	-50.4%

WRAP-UP

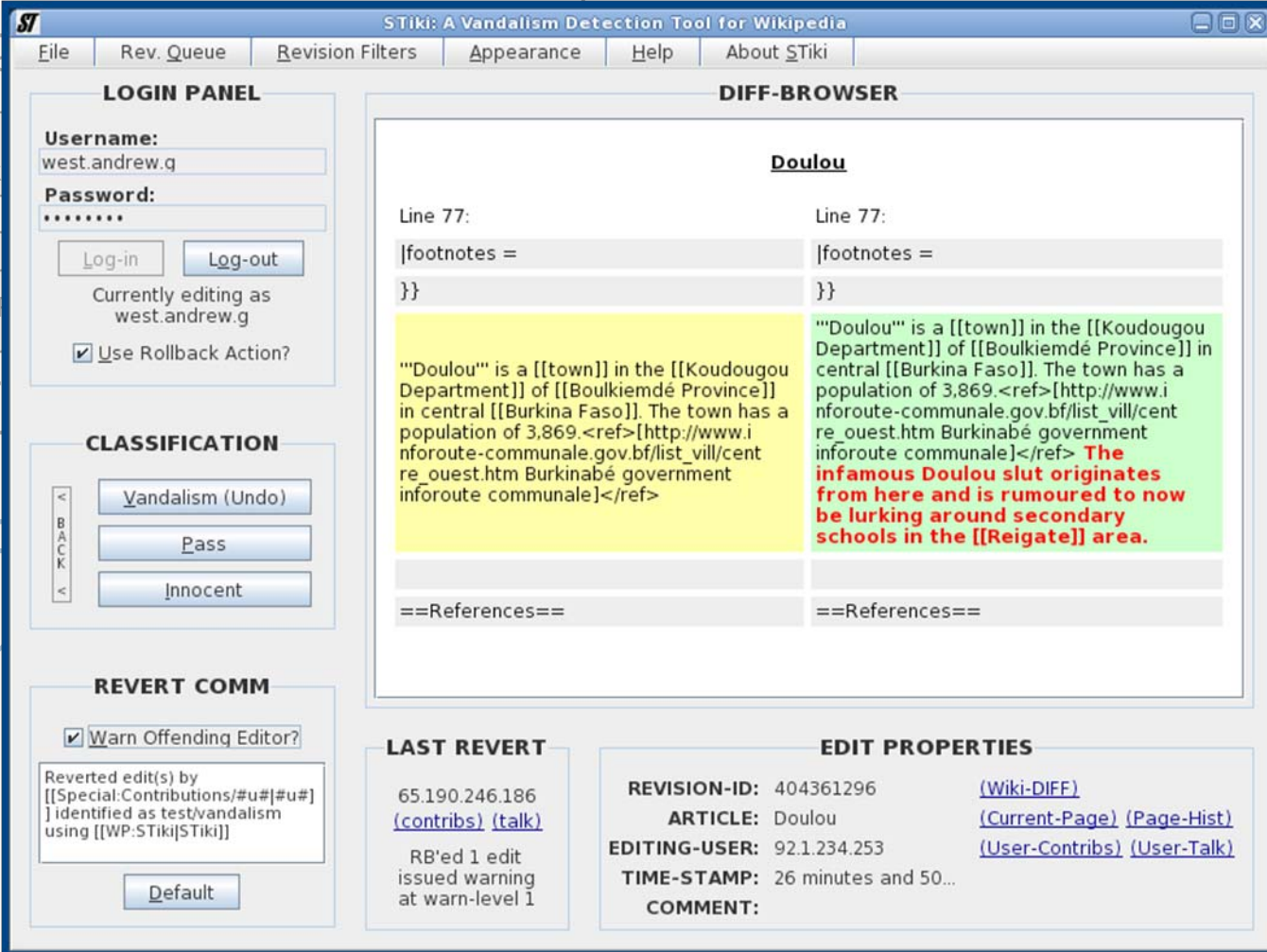
Wikipedia Status Quo



<http://en.wikipedia.org/wiki/WP:STiki>

- Old way: **Brute force** over recent changes
- ClueBot-NG (Bayesian language + metadata)
 - Many reverts; 0.5% FP-tolerance
- STiki **prioritization** for borderline cases
 - Eliminate duplicate human effort

Wikipedia Status Quo



STiki: A Vandalism Detection Tool for Wikipedia

File Rev. Queue Revision Filters Appearance Help About STiki

LOGIN PANEL

Username: west.andrew.g
Password:
Log-in Log-out
Currently editing as west.andrew.g
☒ Use Rollback Action?

CLASSIFICATION

< BACK <
Vandalism (Undo)
Pass
Innocent

REVERT COMM

☒ Warn Offending Editor?
Reverted edit(s) by [[Special:Contributions/#u#/#u#]] identified as test/vandalism using [[WP:STiki|STiki]]
Default

DIFF-BROWSER

Doulou

Line 77: |footnotes =
}}
""Doulou"" is a [[town]] in the [[Koudougou Department]] of [[Boulkiemde Province]] in central [[Burkina Faso]]. The town has a population of 3,869.<ref>[http://www.inforoute-communale.gov.bf/list_vill/centre_ouest.htm Burkinabé government inforoute communale]</ref>
==References==

Line 77: |footnotes =
}}
""Doulou"" is a [[town]] in the [[Koudougou Department]] of [[Boulkiemde Province]] in central [[Burkina Faso]]. The town has a population of 3,869.<ref>[http://www.inforoute-communale.gov.bf/list_vill/centre_ouest.htm Burkinabé government inforoute communale]</ref> **The infamous Doulou slut originates from here and is rumoured to now be lurking around secondary schools in the [[Reigate]] area.**
==References==

LAST REVERT

65.190.246.186
(contribs) (talk)
RB'ed 1 edit issued warning at warn-level 1

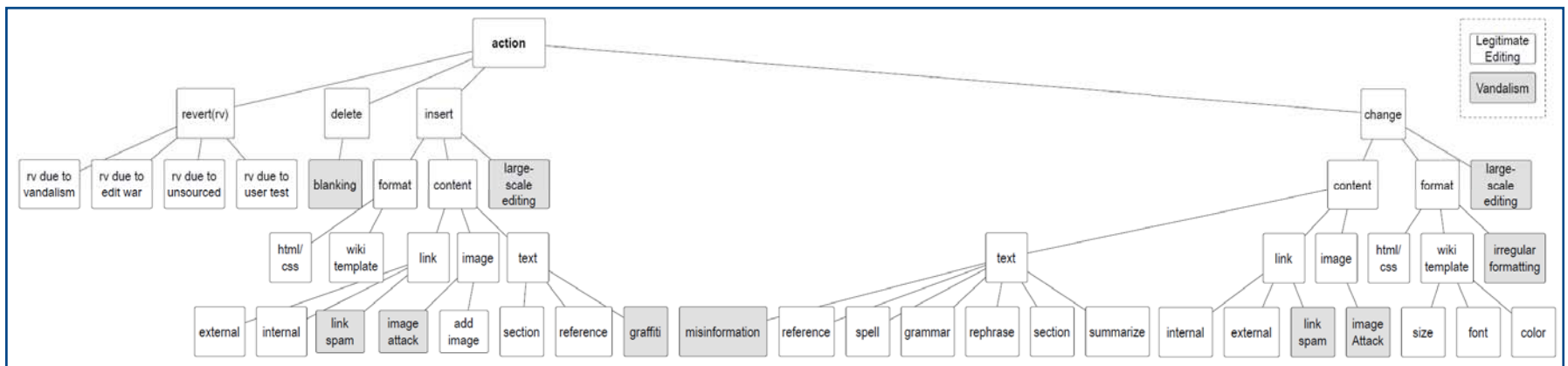
EDIT PROPERTIES

REVISION-ID: 404361296 (Wiki-DIFF)
ARTICLE: Doulou (Current-Page) (Page-Hist)
EDITING-USER: 92.1.234.253 (User-Contribs) (User-Talk)
TIME-STAMP: 26 minutes and 50...
COMMENT:

Future/Ongoing Work



- Bring current technique live
- Classifiers by **vandalism type** [3]
- Concentrating on **acute subsets**
 - Link spam [9] and legally-threatening content
- Other collaborative environments



- Much prior work and benchmarking
- PAN `11: Multiple languages
 - TAKEAWAY: **Language-independent** features sufficient to create well-performing classifier
- PAN `11: Future evidence
 - TAKEAWAY: **Concise and indicative** features, minor performance improvement
- Where to find addl. performance?

- [01] Adler, B.T. and de Alfaro, L.: A content-driven reputation system for the Wikipedia. In: *WWW'07, the World Wide Web Conference*.
- [02] Adler, B., *et al.*: Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In: *CICLing'11 and LNCS*.
- [03] Chin, S., *et al.*: Detecting Wikipedia Vandalism with Active Learning and Statistical Language Models. In *WICOW 2010*.
- [04] Hall, M., *et al.*: The WEKA data mining software: An update. *SIGKDD Explore 11(1)*.
- [05] Potthast, M.: Crowdsourcing a Wikipedia vandalism corpus. In: *SIGIR'10*.
- [06] Potthast, M., *et al.*: Overview of the 1st International competition on Wikipedia vandalism detection. In: *PAN-CLEF 2010 Labs*.
- [07] Priedhorsky, R., *et al.*: Creating, destroying, and restoring value in Wikipedia. In: *ACM GROUP'07, the ACM Conference on Collaborative and Group Work*.
- [08] Velasco, S.M.M.: Wikipedia vandalism detection through machine learning: Feature review and new proposals. *Lab Report for PAN-CLEF 2010*.
- [09] West, A.G., *et al.*: Link Spamming Wikipedia for Profit. In: *WikiSym'11*.
- [10] West, A.G., *et al.*: Detecting Wikipedia vandalism via spatio-temporal analysis of revision metadata. In: *EUROSEC'10, European Workshop on System Security*.

Backup Slides (1)



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

METRIC	GERMAN			ENGLISH			SPANISH		
	RND	ZD	ALL	RND	ZD	ALL	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

- (1) All zero-delay features implemented
- (2) AUCs for Random (RND), Zero-delay (ZD), and ex post facto inclusive (ALL) classifiers

Backup Slides (2)

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_WARNINGS	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.

Backup Slides (3)

	GERMAN	ENGLISH	SPANISH
(a)	1 WT_NO_DELAY	WT_NO_DELAY	USR_EDITS_MONTH
	2 USR_EDITS_EVER	USR_EDITS_MONTH	USR_EDITS_WEEK
	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
(b)	1 NEXT_COMM_VAND (F)	WIKITRUST (F)	NEXT_COMM_VAND (F)
	2 WIKITRUST (F)	WT_DELAY_DELT (F)	NEXT_TIME_AHEAD (F)
	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)”.

	GERMAN	ENGLISH	SPANISH
(a)	1 WT_NO_DELAY	EN_OFFEND_IMPACT	ES_OFFEND_IMPACT
	2 USR_EDITS_MONTH	USR_PG_WARNES	USR_IS_IP
	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
(b)	1 NEXT_COMM_VAND (F)	WIKITRUST (F)	NEXT_COMM_VAND (F)
	2 USR_IS_IP	NEXT_COMM_VAND (F)	USR_EDITS_WEEK
	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.

(left) Features ranked by **info-gain**, (a) without, and (b) with – ex post facto feats.

(left) Best performing **feature subsets** for all language (a) without, and (b) with – ex post facto inclusion