Wiki Vandalysis – Wikipedia Vandalism Analysis

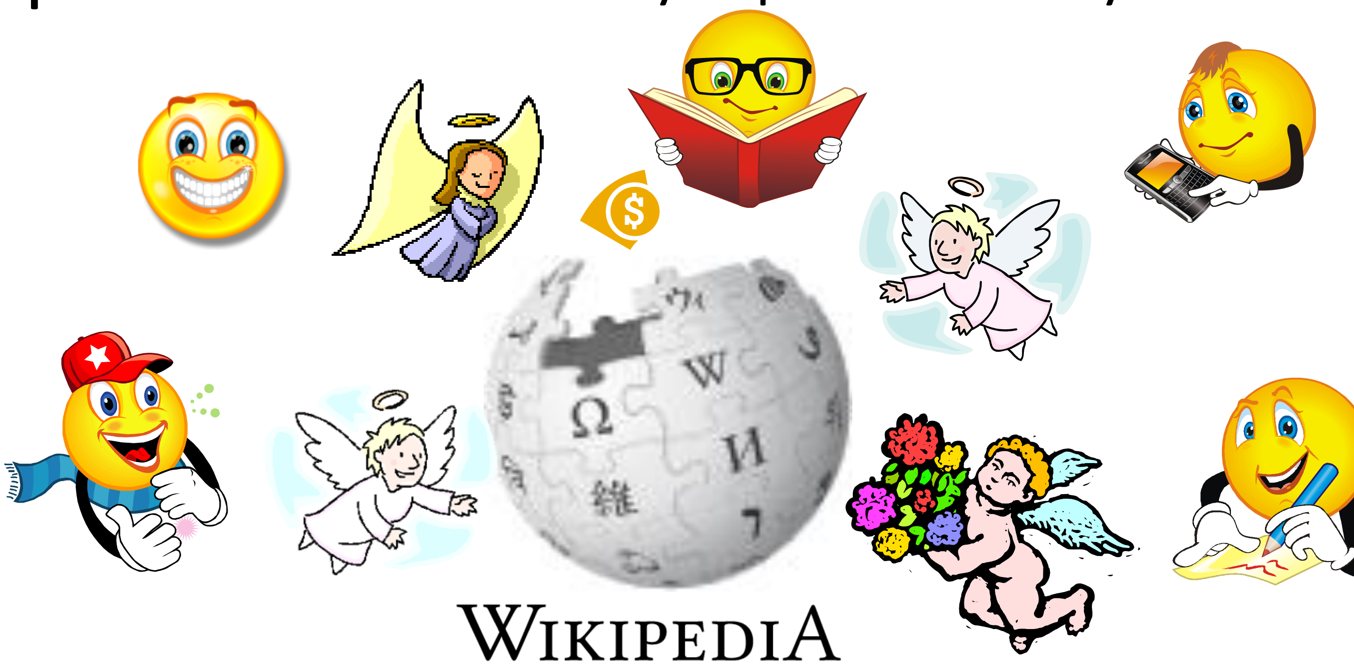
Masters Thesis

by

Manoj Harpalani

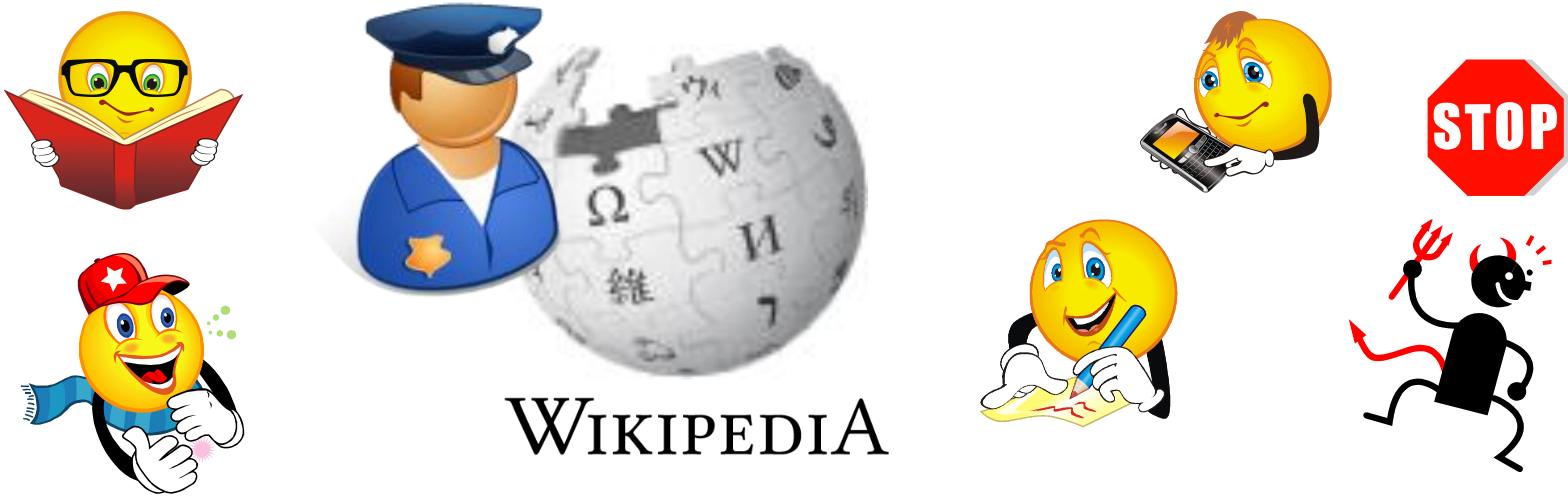
Introduction

- Web 2.0 – Content sharing & collaboration
  - Ex. Social Networking sites, Blogs, Wikis etc
- Wikipedia – “The free encyclopedia that anyone can edit”



Issues with open and free services

Need for Governance & Patrolling



- Publishing policies and ethics are difficult to be expressed as rules
- Rule based policies fall short and are unmanageable
- Need for adaptive and intelligent techniques to detect such attacks
- It’s not just about Wikipedia, Other examples include:
  - Rude comments posted on blogs
  - Bullying on social networking websites
  - Content based sharing policies

Vandalism

- Wikipedia defines Vandalism as “A deliberate attempt to compromise the integrity of Wikipedia”
- Article Integrity refers to :
  - Relevance of edit content
  - Correctness of markup or formatting syntax
  - Stating appropriate and factual information suitable for encyclopedic content

Types of vandalism

- Content Vandalism
  - Silly Vandalism
  - Sneaky Vandalism
- Markup Vandalism
  - Adding irrelevant content in markup
  - Template Vandalism

Silly Vandalism

[http://www.hreonline.com/pdfs/03012008SoftscapeDocu
process of attracting and retaining profitable employees, a
and of strategic importance, has come to be known as "[[tl

+ shut up ugly

+

+

+ and i like poo.

* New York Telephone Company Long Island Headquarters
BellTel Lofts (2006)

* Times Square Building, Rochester, New York (1930)

+

+ he was gay

one time there were over 2000 named "Salvia" species.
700-900 distinct species and subspecies, depending on
18.</ref>

+ like i sayed befor go to hell

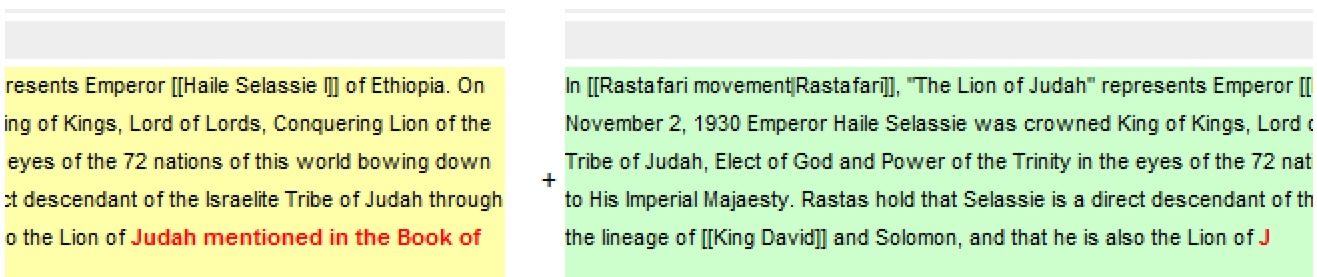
The meaning of the name "Aaron" is unclear.

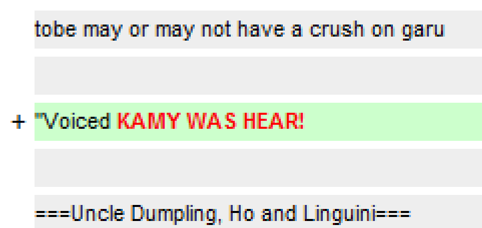
+ #Gay Pregnancy

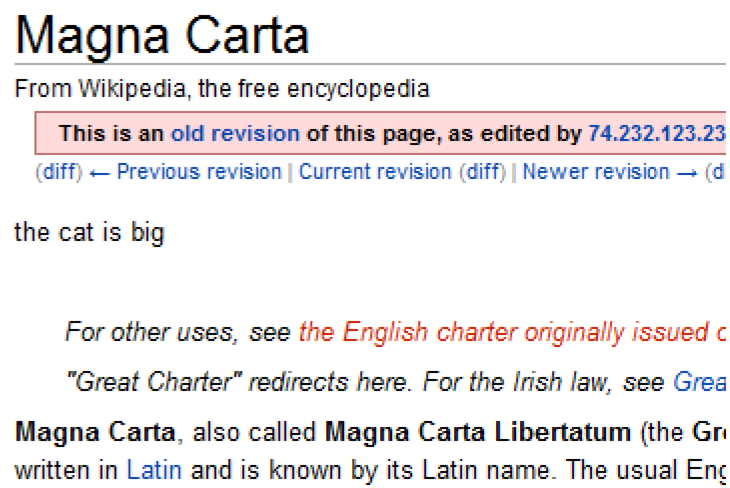
+ # From the faggot

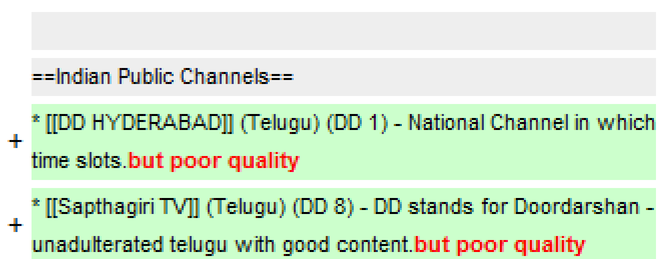
+ # One of gays

Sneaky Vandalism









Markup Vandalism

| [[Beckwourth Pass]] ||{{convert|5221|ft|m|0|abbr=on}} || [[California State Route 70|SR 70]] (paved road)

+ |-theres nothin on this website

| [[Donner Pass]] ||{{convert|7085|ft|m|0|abbr=on}} || [[Interstate 80 in California |I-80]] (interstate highway)

+ [[Category:People who speak with a lisp]]

+ [[Image:California Clipper 500.jpg|right|thumb|uprighhbbbbbbbt=1.5|Sailing to California at the beginning of the Gold Rush]]

Neha is an uncommon first name for women and an equally uncomm
(1990 U.S. Census)

+ [[File:C:\Documents and Settings\PU00126\Desktop\a.jpg]]

Template Vandalism

Revision as of 10:54, 20 November 2009

Square kilometre (U.S. spelling: **square kilometer**), symbol **km^2**, is a decimal multiple of the SI unit of surface are

- 1,000,000 m^2
- 100 ha (hectare)
- 0.386102 square miles
- 247.105381 acres

> ***This article may meet Wikipedia's criteria for speedy deletion*, but no reason has been given for**
> *of the speedy deletion criteria. Replace this tag with `{{db|1=some reason}}`.*
>
> If this article does not meet the criteria for speedy deletion, or you intend to fix it, please remove this
> **created yourself**. If you created this page and you disagree with its proposed speedy deletion, plea

John Thain

From Wikipedia, the free encyclopedia

This is an old revision of this page, as edited by 64.201.173.145 (talk) at 17:21, 30 November 2009. It may differ significantl

(diff) ← Previous revision | Current revision (diff) | Newer revision → (diff)

{{Infobox Person

Hey Thatguyflint, back off ya bum. you dont own this web page. my comments are constructive and well researched.

|name = John Thain |image = John Thain briefing.jpg |image_size = |caption = Thain in 2006. |birth_name = John Alex
|death_place = |death_cause = |nationality = |other_names = |known_for = |education = |alma_mater = |employer = |o
million[1] |networth = |term = |predecessor = |successor = |party = |boards = INSEAD, MIT Sloan School of Managem

John Alexander Thain (born May 26, 1955) is an American businessman and investment banker.

Prior approaches

- Rule Based
  - ClueBot
- Machine Learning :
  - Naïve Bayes – Bag of words
  - Compression ratio
  - PAN 2010 Workshop
- Natural Language Processing :
  - “Got you!” Vandalism Detection using Shallow Syntactic and Semantic modeling.

Our contribution

- PAN 2010 Workshop - Introduce informative features
  - Our results: AUC 88.5 %
  - Winner results : AUC 92%
- Improve on our features from the learning of PAN proceedings.
- Introduce a new approach inspired from Authorship Attribution using PCFG to model the syntax and style.
- Analyze impact of balanced and unbalanced dataset on results.
- Compare our performance with the Syntax and Semantic approach by “Got you!” study.
- Analyze the performance of our classifier for each edit type insert or change, delete and template edits individually.

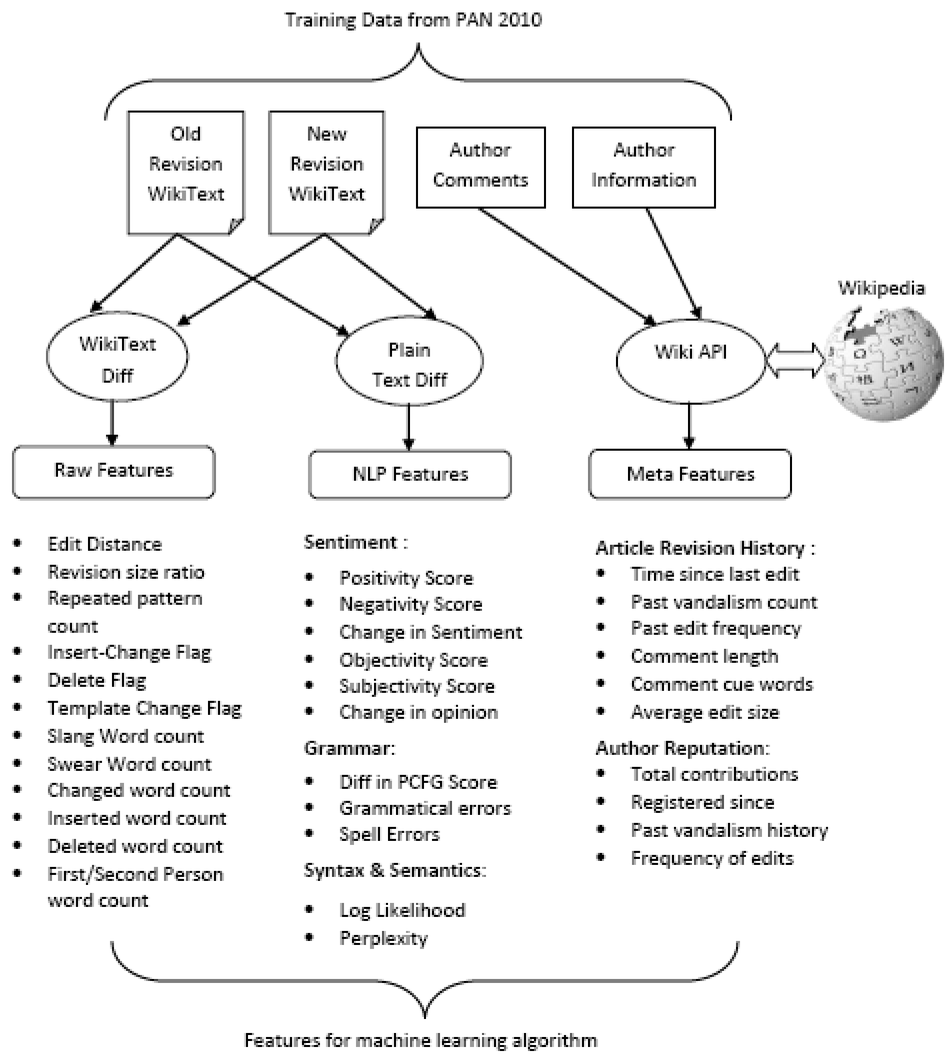
Problem Definition

- Given an edit in Wikipedia, we can use the below information for the vandalism classification task :
  - The edit itself
  - Previous contributions of the editor
  - Comments of the edit
  - Past revisions of the edit
  - Related articles from the web or in Wikipedia itself

Edit Types

- Content changes
  - Insert – Addition of new content
  - Change – Modification of existing content
  - Delete – Removal of existing content
- Wiki Markup changes
  - Short change in visible content
  - Change in formatting/styling
  - Insertion of links or images

Feature Extraction



Features – a closer look

Raw Features

- Edit Distance
- Revision size ratio
- Repeated pattern count
- Insert-Change Flag
- Delete Flag
- Template Change Flag
- Slang Word count
- Swear Word count
- Changed word count
- Inserted word count
- Deleted word count
- First/Second Person word count

NLP Features

Sentiment :

- Positivity Score
- Negativity Score
- Change in Sentiment
- Objectivity Score
- Subjectivity Score
- Change in opinion

Grammar:

- Diff in PCFG Score
- Grammatical errors
- Spell Errors

Syntax & Semantics:

- Log Likelihood
- Perplexity

Meta Features

Article Revision History :

- Time since last edit
- Past vandalism count
- Past edit frequency
- Comment length
- Comment cue words
- Average edit size

Author Reputation:

- Total contributions
- Registered since
- Past vandalism history
- Frequency of edits

Features for machine learning algorithm

Features – a closer look

Raw Features

- Edit Distance
- Revision size ratio
- Repeated pattern count
- Insert-Change Flag
- Delete Flag
- Template Change Flag
- Slang Word count
- Swear Word count
- Changed word count
- Inserted word count
- Deleted word count
- First/Second Person word count

NLP Features

Sentiment :

- Positivity Score
- Negativity Score
- Change in Sentiment
- Objectivity Score
- Subjectivity Score
- Change in opinion

Grammar:

- Diff in PCFG Score
- Grammatical errors
- Spell Errors

Syntax & Semantics:

- Log Likelihood
- Perplexity

Meta Features

Article Revision History :

- Time since last edit
- Past vandalism count
- Past edit frequency
- Comment length
- Comment cue words
- Average edit size

Author Reputation:

- Total contributions
- Registered since
- Past vandalism history
- Frequency of edits

Features for machine learning algorithm

Features – a closer look

Raw Features

- Edit Distance
- Revision size ratio
- Repeated pattern count
- Insert-Change Flag
- Delete Flag
- Template Change Flag
- Slang Word count
- Swear Word count
- Changed word count
- Inserted word count
- Deleted word count
- First/Second Person word count

NLP Features

Sentiment :

- Positivity Score
- Negativity Score
- Change in Sentiment
- Objectivity Score
- Subjectivity Score
- Change in opinion

Grammar:

- Diff in PCFG Score
- Grammatical errors
- Spell Errors

Syntax & Semantics:

- Log Likelihood
- Perplexity

Meta Features

Article Revision History :

- Time since last edit
- Past vandalism count
- Past edit frequency
- Comment length
- Comment cue words
- Average edit size

Author Reputation:

- Total contributions
- Registered since
- Past vandalism history
- Frequency of edits

Features for machine learning algorithm

Features – a closer look

Raw Features

- Edit Distance
- Revision size ratio
- Repeated pattern count
- Insert-Change Flag
- Delete Flag
- Template Change Flag
- Slang Word count
- Swear Word count
- Changed word count
- Inserted word count
- Deleted word count
- First/Second Person word count

NLP Features

Sentiment :

- Positivity Score
- Negativity Score
- Change in Sentiment
- Objectivity Score
- Subjectivity Score
- Change in opinion

Grammar:

- Diff in PCFG Score
- Grammatical errors
- Spell Errors

Syntax & Semantics:

- Log Likelihood
- Perplexity

Meta Features

Article Revision History :

- Time since last edit
- Past vandalism count
- Past edit frequency
- Comment length
- Comment cue words
- Average edit size

Author Reputation:

- Total contributions
- Registered since
- Past vandalism history
- Frequency of edits

Features for machine learning algorithm

Features – a closer look

Raw Features

- Edit Distance
- Revision size ratio
- Repeated pattern count
- Insert-Change Flag
- Delete Flag
- Template Change Flag
- Slang Word count
- Swear Word count
- Changed word count
- Inserted word count
- Deleted word count
- First/Second Person word count

NLP Features

Sentiment :

- Positivity Score
- Negativity Score
- Change in Sentiment
- Objectivity Score
- Subjectivity Score
- Change in opinion

Grammar:

- Diff in PCFG Score
- Grammatical errors
- Spell Errors

Syntax & Semantics:

- Log Likelihood
- Perplexity

Meta Features

Article Revision History :

- Time since last edit
- Past vandalism count
- Past edit frequency
- Comment length
- Comment cue words
- Average edit size

Author Reputation:

- Total contributions
- Registered since
- Past vandalism history
- Frequency of edits

Features for machine learning algorithm

Features – a closer look

Raw Features

- Edit Distance
- Revision size ratio
- Repeated pattern count
- Insert-Change Flag
- Delete Flag
- Template Change Flag
- Slang Word count
- Swear Word count
- Changed word count
- Inserted word count
- Deleted word count
- First/Second Person word count

NLP Features

Sentiment :

- Positivity Score
- Negativity Score
- Change in Sentiment
- Objectivity Score
- Subjectivity Score
- Change in opinion

Grammar:

- Diff in PCFG Score
- Grammatical errors
- Spell Errors

Syntax & Semantics:

- Log Likelihood
- Perplexity

Meta Features

Article Revision History :

- Time since last edit
- Past vandalism count
- Past edit frequency
- Comment length
- Comment cue words
- Average edit size

Author Reputation:

- Total contributions
- Registered since
- Past vandalism history
- Frequency of edits

Features for machine learning algorithm

Sentiment Analysis

- Expressing personal opinions and negative facts is common in many vandalism edits in Wikipedia.
- LingPipe’s Sentiment Analysis Tool
  - # of subjective and objective sentences
  - # of positive and negative sentences
  - Change in positivity and negativity score
  - Change in objectivity and subjectivity score

Grammar

- Vandals have a different writing style and syntax than regular contributors
- Model the syntax and style of regular editors and vandals
  - Regular Sentence Parser trained only on regular edits
  - Vandalism Sentence Parser trained only on vandalism edits
- Compute the log probability (PCFG score) of the best parse from the trained parser.
- For each edit compute statistics like min, max, mean, sum and standard deviation from the PCFG score of all sentences.
- Calculate the diff between the statistics from regular and vandalism parser to use it as a feature

Syntactic and Semantic Modeling

- Large number of vandalisms are off topic
- Tricky to be captured without additional information
- Re-implement “Got you!” Vandalism Detection using Shallow Syntactic and Semantic modeling

- For each edit
  - Get top 100 search results from Bing
  - Build tri-gram language model for each article on :
    - Unigram & POS Tags to capture semantics.
    - Only POS tags to capture syntax.
  - Calculate the log likelihood and perplexity of the edit diff on the trained tri-gram language models

Re-implementing "Got you!" Syntax & Semantic Modeling

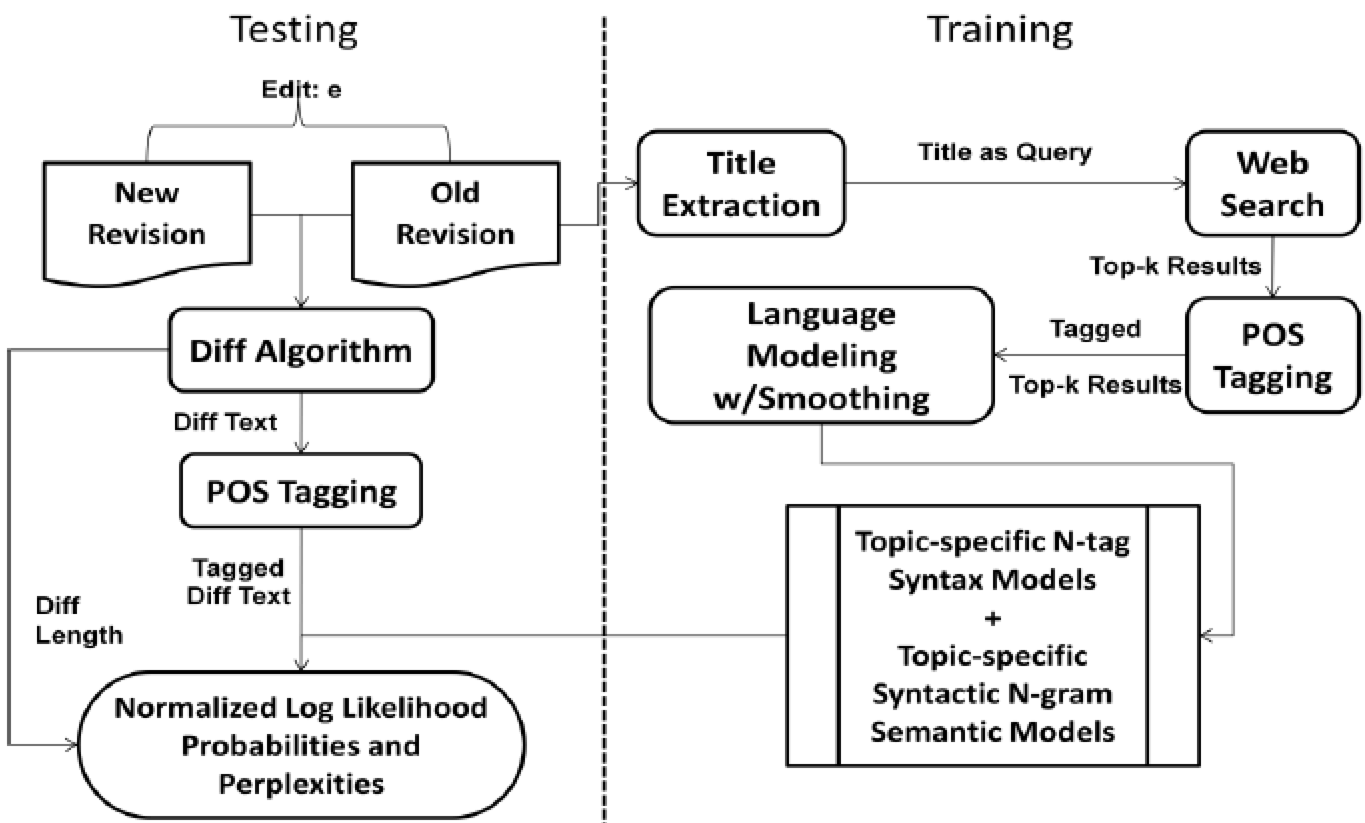


Figure 1. Topic-specific N-tag Syntax Models and Syntactical N-gram for Syntactical and Semantic Modeling

Classifiers

- Experimented with various classifiers :
  - C4.5 decision trees
  - AdaBoost
  - SVM
  - Naive Bayes Tree
  - LogitBoost
- LogitBoost a boosting technique combined with a logistic regression classifier performed the best among all and achieved an AUC of 94% with F-Measure of 53% with 10 fold cross validation.

Evaluation Overview

- Corpus:
  - PAN 2010 Workshop
  - 32,444 human annotated edits - 2904 vandalisms
- Balanced v/s Unbalanced dataset
  - Vandalism to Regular ratio 1:10
  - “Got you!” Syntax & Semantics – Balanced
  - PAN 2010 Workshop – Unbalanced
  - We evaluate our classifier on both
- Baseline
  - PAN Workshop - Unbalanced dataset
  - “Got you!” study - Balanced dataset

Evaluation Metrics

- Accuracy More than 90% Easy!
  - Just output Regular
- Precision

  # Actual vandalisms Identified

  # Vandalisms Identified

- Recall

  # Actual vandalism identified

  Total # of actual vandalisms

- F1 – Harmonic mean of precision and recall
- AUC – True positive v/s false positive rate

Evaluation & Results on Unbalanced Dataset

- Complete PAN 2010 corpus(Unbalanced):
  - Total corpus: 32444
  - Training corpus: 15000
  - Test corpus: 17444
  - Ratio of vandalism to regular -> 1:10

Experiment	Precision	Recall	F–Measure	AUC
PAN 2010 Winner	0.86	0.56	0.67	0.92
Our PAN 2010 setting results	0.64	0.35	0.45	0.91
10 fold cross validation on complete PAN 2010 corpus	0.74	0.41	0.53	0.94

* Without PCFG & Syntax Semantics

Evaluation & Results on Balanced Dataset

- Syntax & Semantics:
  - Balanced Corpus:
    - Random Sampling
    - Equal # of regular & vandalism edits
    - Features on Inserts and Changes

Experiment	Training size
“Got you!”	1600
Syntax & Semantics w/o PCFG	4036
Syntax & Semantics w/ PCFG	4036

Experiment	Precision	Recall	F–Measure	AUC
”Got you!” [Wang and McKeown, 2010]	0.85	0.85	0.86	–
Our features with Syntax and Semantics without PCFG	0.83	0.89	0.86	0.93
Our features with Syntax and Semantics with PCFG	0.84	0.89	0.87	0.94

Unbalanced v/s Balanced

- Unbalanced Corpus:
  - Training size : 4036 edits, 495 vandalisms

Experiment	Precision	Recall	F–Measure	AUC
Our features without Syntax and Semantics	0.74	0.43	0.54	0.91
Our features with Syntax and Semantics	0.74	0.48	0.58	0.92

– Balanced Corpus:

Experiment	Precision	Recall	F–Measure	AUC
"Got you!" [Wang and McKeown, 2010]	0.85	0.85	0.86	–
Our features with Syntax and Semantics without PCFG	0.83	0.89	0.86	0.93
Our features with Syntax and Semantics with PCFG	0.84	0.89	0.87	0.94

Classification on Edit Types

- Vandalism breakup by edit type :
  - Insert or change 80% , Template change 17%, Delete 3%

Experiment	Training size
Insert or change w/ PCFG	10086
Insert or change w/o PCFG	10086
Delete w/o PCFG	3000
Template changes	13000

Experiment	Precision	Recall	F–Measure	AUC
Insert or Changes without PCFG	0.73	0.41	0.52	0.92
Insert or Change with PCFG	0.73	0.48	0.58	0.93
Delete without PCFG	0.58	0.25	0.35	0.95
Template Change edits	0.71	0.14	0.23	0.93

Top 10 Features for Insert or Changes

Feature	Information Gain
Total number of author contributions	0.105
How long the author has been registered	0.098
How frequently the author contributed in the training set	0.097
If the author is a registered user	0.088
Maximum PCFG Score Difference	0.043
How often the article has been reverted	0.037
Total contributions of author to Wikipedia	0.034
Previous vandalism count of the article	0.032
Length of edit comment	0.029
Revision Size Ratio	0.024

Top 10 Features for Deletes

Feature	Information Gain
Revision Size Ratio	0.027
Edit Distance	0.026
Deleted Word Count	0.023
Total Author contributions in Wikipedia	0.022
Total sentences deleted in	0.018
No. of objective sentences deleted	0.016
Is the author registered on Wikipedia?	0.015
How long the author has been registered	0.014
No. of subjective sentences deleted	0.013
Comment Length	0.009

Top 10 Features for Template changes

Feature	Information Gain
Total contributions of author to Wikipedia	0.044
How long the author has been registered	0.035
If the author is a registered user	0.032
How frequently the author contributed in the training set	0.025
How many times the article has been reverted previously	0.016
How many revisions have been made previously for the article	0.015
How many times the articles has been vandalized in the past	0.015
Average number of edits per month	0.014
Comment Length	0.014
Average time beteween edits	0.012

Thank you

- Prof. Rob Johnson to guide and motivate me to take up this challenging project.
- Prof. Luis Ortiz, Prof. Tamara Berg and Prof. Yejin Choi for their suggestions and advise.
- Michael, Thanadit, Megha & Sandesh for their valuable contributions in the project.
- Team ! Would not have been possible without you guys.
- Thanks to the wonderful audience.

Questions & Answers

?