Wiki Vandalysis- Wikipedia Vandalism Analysis

Manoj Harpalani, Thanadit Phumprao, Megha Bassi, Michael Hart, and Rob Johnson

Stony Brook University

Text Features

- Edit Distance
- Text Changes
- Spelling Errors
- Obscene Words
- Repeated Patterns
- Sum of metrics
  - Spelling errors, obscene words, repeated patterns
- Sentences inserted, deleted and changed
- Word count
- Ratio of suspicious features to the article word count.

Advanced Text Analysis Features

- Grammar
  - Link grammar checker
  - Discover number of grammatical errors.
- Sentiment Analysis
  - Logistic regression over character-level n-grams
  - Trained on film summaries and reviews
  - Measure both polarity and subjectivity
    - Across edit type (insert,delete,modify)
    - Across sentences
    - Over all words

Meta-Features

- Article
  - Number of times article was vandalized previously
  - Number times article was reverted previously
- Editor
  - Time since author registered in Wikipedia
  - Number of previous vandalisms
  - Total contributions to Wikipedia
  - Total contributions to a given article
  - Number of contributions in a sampling of edits

Classification approaches

- Baseline
  - Used Bag of Words approach
  - Added RankBoost to improve baseline

- Classifiers built on features
  - Naive Bayes
  - C4.5 Decision Tree
  - NBTree

Classifiers evaluated

Evaluation Results on Training Set:

Metric	NB+BoW	NB+BoW+RankBoost	NB	C4.5	NBTree
Precision	27.8%	34.1%	15.8%	53.2%	**64.3%**
Recall	32.6%	26.6%	**93.2%**	36.9%	36.4%
Accuracy	87.5%	89.7%	69.2%	94.1%	**94.8%**
F-measure	30.1%	29.9%	27.1%	43.6%	**46.5%**
AUC	69%	62%	88.5%	80.5%	**91%**

Evalutation Results on Test Set:

Metric	NB	C4.5	NBTree
Precision	19.0%	51.0%	**61.5%**
Recall	**92.0%**	26.7%	25.2%
Accuracy	72.0%	91.6%	**92.3%**
F-measure	35.5%	35.1%	**35.8%**
AUC	86.6%	76.9%	**88.7%**

Performance for Selected users

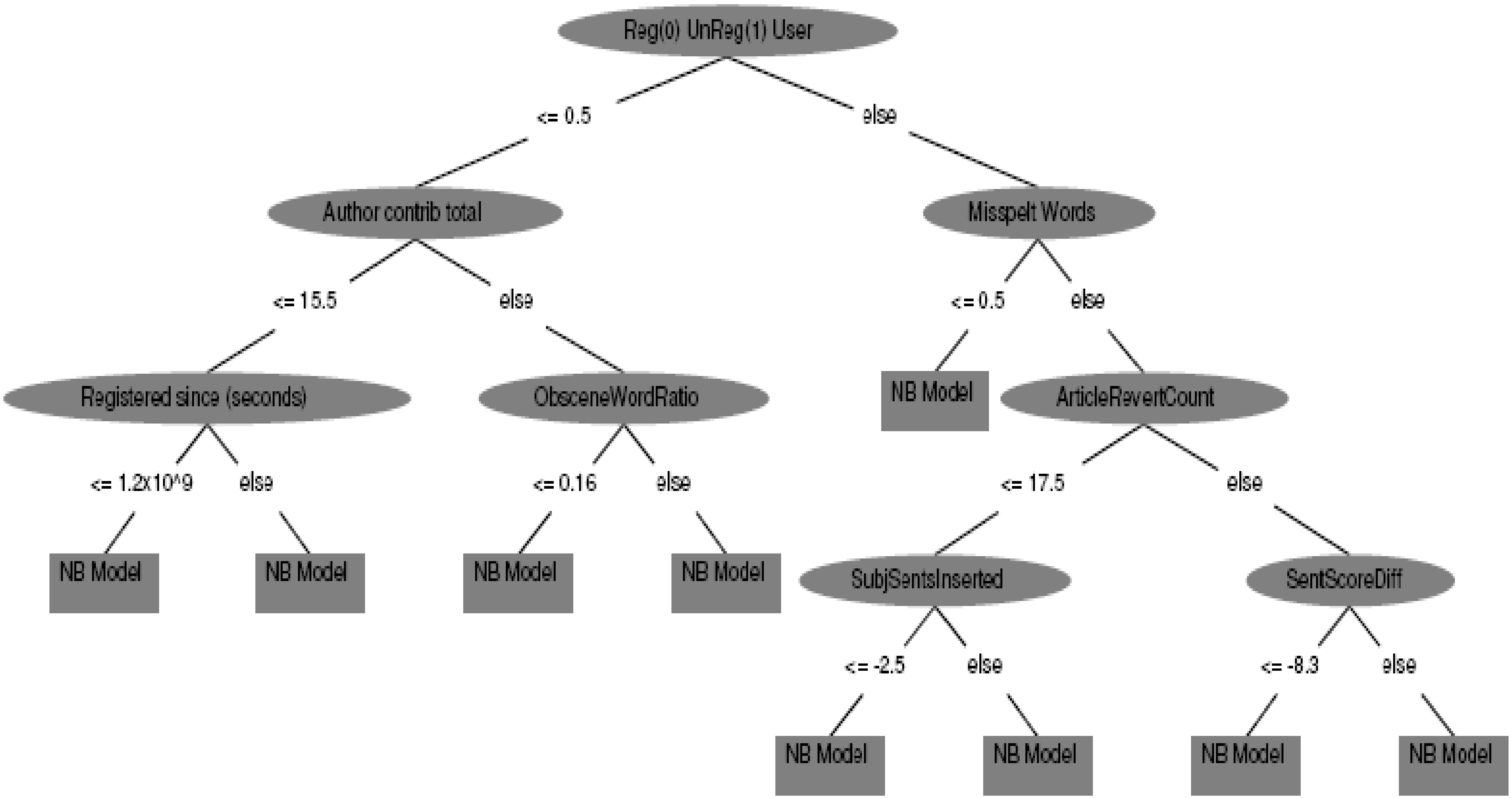
Type of user	FP rate	Recall	Precision
Registered users	< 0.1%	22.0%	68.4%
Registered users that edited this article 10 times or more	< 0.01%	0.0%	0.0%
Unregistered users	3.9%	40.8%	67.2%
IP addresses that edited this article 10 times or more	1.7%	33.3%	50.0%

Top Performing Features

Feature	Information Gain
Total number of author contributions	0.074
How long the author has been registered	0.067
If the author is a registered user	0.06
How frequently the author contributed in the training sex	0.04
How often the article has been vandalized	0.035
How often the article has been reverted	0.034
The number of previous contributions on the article	0.019
Change in sentiment score	0.019
Number of misspelled words	0.019
Sum of metrics	0.018

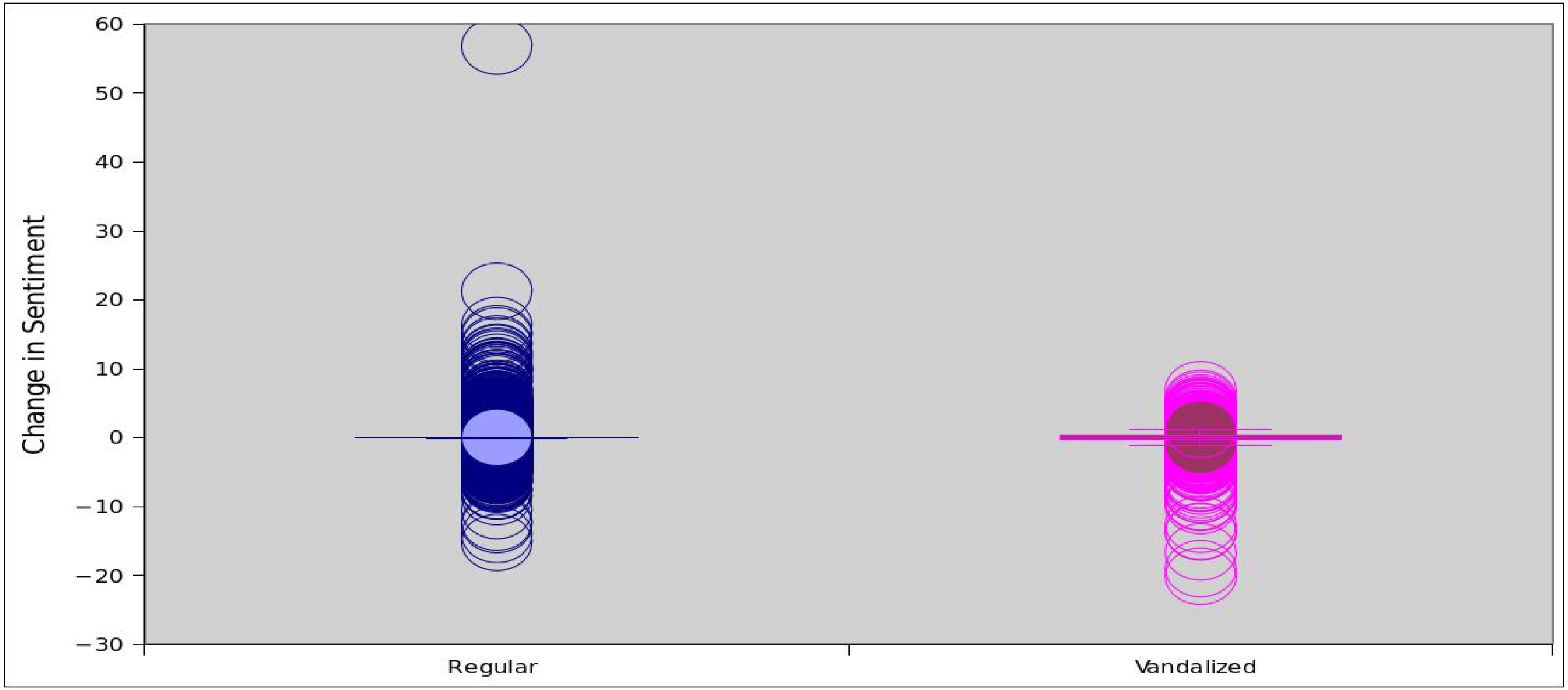
Meta feature
Text feature
Advanced text feature

Features Employed by the NBTree



Sentiment and Vandalism

- Change in polarity and vandalism
  - Vandalism skewed negatively
  - Regular edits skewed positively
- 0:03 with a standard deviation of 1:1



Timely suggestions for Wikipedia

- Certain IPs contribute heavily to Wikipedia
  - IPs belong to universities, Redmond, etc.
  - Recruit!

- Incorporate simple features into current vandalism tools
  - Editor meta-information
  - Article meta-information
  - Even if not used directly to classify vandalism
    - Use to rank suspicious edits for Wiki Admins

Vandalism of Registered Users is **hard**

- Our classifier strengths
  - Unregistered users
  - IPs that contribute frequently
  - Registered users with minimal site usage

- But poor classification of active registered users
  - Not many instances of vandalism by these users
    - Our features provide little discriminatory information
  - Vandalism not as clear-cut

- Suggestions
  - Ignore? Apply the Law of Diminishing returns ☺
  - Use techniques from imbalanced training set

Conclusions

- NBTree worked well by partitioning edits
  - Train a tailored stochastic model
  - Suggests a one-size fits all approach is difficult
    - Until someone creates a better model describing vandalism
- Author and article meta information incredibly useful
  - Expectation of the quality of the edit
- Main limitation
  - Could not verify relevance/factuality of content
  - Ideas?
    - Expertise of editor
    - Language model based on similar articles
    - Value-added assessment

Grazie! Domande?