Language of Vandalism: Improving Wikipedia Vandalism Detection via Stylometric Analysis

Manoj Harpalani, Michael Hart, Sandesh Singh, Rob Johnson, and Yejin Choi
Department of Computer Science
Stony Brook University
NY 11794, USA
{mharpalani, mhart, sssingh, rob, ychoi}@cs.stonybrook.edu

Abstract

Community-based knowledge forums, such as Wikipedia, are susceptible to *vandalism*, i.e., ill-intentioned contributions that are detrimental to the quality of collective intelligence. Most previous work to date relies on shallow lexico-syntactic patterns and metadata to automatically detect vandalism in Wikipedia. In this paper, we explore more linguistically motivated approaches to vandalism detection. In particular, we hypothesize that textual vandalism constitutes a unique *genre* where a group of people share a similar linguistic behavior. Experimental results suggest that (1) statistical models give evidence to unique language styles in vandalism, and that (2) deep syntactic patterns based on probabilistic context free grammars (PCFG) discriminate vandalism more effectively than shallow lexico-syntactic patterns based on n-grams.

1 Introduction

Wikipedia, the "free encyclopedia" (Wikipedia, 2011), ranks among the top 200 most visited websites worldwide (Alexa, 2011). This editable encyclopedia has amassed over 15 million articles across hundreds of languages. The English language encyclopedia alone has over 3.5 million articles and receives over 1.25 million edits (and sometimes upwards of 3 million) daily (Wikipedia, 2010). But allowing anonymous edits is a double-edged sword; nearly 7% (Potthast, 2010) of edits are vandalism, i.e. revisions to articles that undermine the quality and veracity of the content. As Wikipedia continues to grow, it will become increasingly infeasible for Wikipedia users and administrators to manually police articles. This pressing issue has spawned recent research activities to understand and counteract vandalism (e.g., Geiger and Ribes (2010)). Much of previous work relies on hand-picked rules such as lexical cues (e.g., vulgar words) and metadata (e.g., anonymity, edit frequency) to automatically detect vandalism in Wikipedia (e.g., Potthast et al. (2008), West et al. (2010)). Although some recent work has started exploring the use of natural language processing, most work to date is based on shallow lexico-syntactic patterns (e.g., Wang and McKeown (2010), Chin et al. (2010), Adler et al. (2011)).

We explore more linguistically motivated approaches to detect vandalism in this paper. Our hypothesis is that textual vandalism constitutes a unique *genre* where a group of people share similar linguistic behavior. Some obvious hallmarks of this style include usage of obscenities, misspellings, and slang usage, but we aim to automatically uncover stylistic cues to effectively discriminate between vandalizing and normal text. Experimental results suggest that (1) statistical models give evidence to unique language styles in vandalism, and that (2) deep syntactic patterns based on probabilistic context free grammar (PCFG) discriminate vandalism more effectively than shallow lexico-syntactic patterns based on n-grams.

2 Stylometric Features

Stylometric features attempt to recognize patterns of style in text. These techniques have been traditionally applied to attribute authorship (Argamon et al. (2009), Stamatatos (2009)), opinion mining

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 83–88,
Portland, Oregon, June 19-24, 2011.

(Panicheva et al., 2010), and forensic linguistics (Turell, 2010). For our purposes, we hypothesize that different stylistic features appear in regular and vandalizing edits. For regular edits, honest editors will strive to follow the stylistic guidelines set forth by Wikipedia (e.g. objectivity, neutrality and factuality). For edits that vandalize articles, these users may converge on common ways of vandalizing articles.

2.1 Language Models

To differentiate between the styles of normal users and vandalizers, we employ language models to capture the stylistic differences between authentic and vandalizing revisions. We train two trigram language model (LM) with Good-Turing discounting and Katz backoff for smoothing of vandalizing edits (based on the text difference between the vandalizing and previous revision) and good edits (based on the text difference between the new and previous revision).

2.2 Probabilistic Context Free Grammar (PCFG) Models

Probabilistic context-free grammars (PCFG) capture deep syntactic regularities beyond shallow lexico-syntactic patterns. Raghavan et al. (2010) reported for the first time that PCFG models are effective in learning stylometric signature of authorship at deep syntactic levels. In this work, we explore the use of PCFG models for vandalism detection, by viewing the task as a genre detection problem, where a group of authors share similar linguistic behavior. We give a concise description of the use of PCFG models below, referring to Raghavan et al. (2010) for more details.

(1) Given a training corpus D for vandalism detection and a generic PCFG parser C_o trained on a manually tree-banked corpus such as WSJ or Brown, tree-bank each training document $d_i \in D$ using the generic PCFG parser C_o.

(2) Learn vandalism language by training a new PCFG parser C_{vandal} using only those tree-banked documents in D that correspond to vandalism. Likewise, learn regular Wikipedia language by training a new PCFG parser $C_{regular}$ using only those tree-banked documents in D that correspond to regular Wikipedia edits.

(3) For each test document, compare the probability of the edit determined by C_{vandal} and $C_{regular}$, where the parser with the higher score determines the class of the edit.

We use the PCFG implementation of Klein and Manning (2003).

3 System Description

Our system decides if an edit to an article is vandalism by training a classifier based on a set of features derived from many different aspects of the edit. For this task, we use an annotated corpus (Potthast et al., 2010) of Wikipedia edits where revisions are labeled as either vandalizing or non-vandalizing. This section will describe in brief the features used by our classifier, a more exhaustive description of our non-linguistically motivated features can be found in Harpalani et al. (2010).

3.1 Features Based on Metadata

Our classifier takes into account metadata generated by the revision. We generate features based on author reputation by recording if the edit is submitted by an anonymous user or a registered user. If the author is registered, we record how long he has been registered, how many times he has previously vandalized Wikipedia, and how frequent he edits articles. We also take into account the comment left by an author. We generate features based on the characteristics of the articles revision history. This includes how many times the article has been previously vandalized, the last time it was edited, how many times it has been reverted and other related features.

3.2 Features Based on Lexical Cues

Our classifier also employs a subset of features that rely on lexical cues. Simple strategies such as counting the number of vulgarities present in the revision are effective to capture obvious forms of vandalism. We measure the edit distance between the old and new revision, the number of repeated patterns, slang words, vulgarities and pronouns, the type of edit (insert, modification or delete) and other similar features.

Features	P	R	F1	AUC
Baseline	72.8	41.1	52.6	91.6
+LM	73.3	42.1	53.5	91.7
+PCFG	**73.5**	**47.7**	**57.9**	92.9
+LM+PCFG	73.2	47.3	57.5	**93.0**

Table 1: Results on naturally unbalanced test data

3.3 Features Based on Sentiment

Wikipedia editors strive to maintain a neutral and objective voice in articles. Vandals, however, insert subjective and polar statements into articles. We build two classifiers based on the work of Pang and Lee (2004) to measure the polarity and objectivity of article edits. We train the classifier on how many positive and negative sentences were inserted as well as the overall change in the sentiment score from the previous version to the new revision and the number of inserted or deleted subjective sentences in the revision.

3.4 Features Based on Stylometric Measures

We encode the output of the LM and PCFG in the following manner for training our classifier. We take the log-likelihood of the regular edit and vandalizing edit LMs. For our PCFG, we take the difference between the minimum log-likelihood score (i.e. the sentences with the minimum log-likelihood) of C_{vandal} and $C_{regular}$, the difference in the maximum log-likelihood score, the difference in the mean log-likelihood score, the difference in the standard deviation of the mean log-likelihood score and the difference in the sum of the log-likelihood scores.

3.5 Choice of Classifier

We use Weka's (Hall et al., 2009) implementation of LogitBoost (Friedman et al., 2000) to perform the classification task. We use Decision Stumps (Ai and Langley, 1992) as the base learner and run LogitBoost for 500 iterations. We also discretize the training data using the Multi-Level Discretization technique (Perner and Trautzsch, 1998).

4 Experimental Results

Data We use the 2010 PAN Wikipedia vandalism corpus Potthast et al. (2010) to quantify the benefit of stylometric analysis to vandalism detection. This corpus comprises of 32452 edits on 28468 articles, with 2391 of the edits identified as vandalism by human annotators. The class distribution is highly skewed, as only 7% of edits corresponds to vandalism. Among the different types of vandalism (e.g. deletions, template changes), we focus only on those edits that inserted or modified text (17145 edits in total) since stylometric features are not relevant to deletes and template modifications. Note that insertions and modifications are the main source for vandalism.

Feature	Score
Total number of author contributions	0.106
How long the author has been registered	0.098
How frequently the author contributed in the training set	0.097
If the author is registered	0.0885
Difference in the maximum PCFG scores	0.0437
Difference in the mean PCFG scores	0.0377
How many times the article has been reverted	0.0372
Total contributions of author to Wikipedia	0.0343
Previous vandalism count of the article	0.0325
Difference in the sum of PCFG scores	0.0320

Table 2: Top 10 ranked features on the unbalanced test data by InfoGain

We randomly separated 15000 edits for training of C_{vandal} and $C_{regular}$, and 17444 edits for testing, preserving the ratio of vandalism to non-vandalism revisions. We eliminated 7359 of the testing edits to remove revisions that were exclusively template modifications (e.g. inserting a link) and maintain the observed ratio of vandalism for a total of 10085 edits. For each edit in the test set, we compute the probability of each modified sentence for C_{vandal} and $C_{regular}$ and generate the statistics for the features described in 3.4. We compare the performance of the language models and stylometric features against a baseline classifier that is trained on metadata, lexical and sentiment features using 10 fold stratified cross validation on the test set.

Results Table 1 shows the experimental results. Because our dataset is highly skewed (97% corresponds to *"not vandalism"*), we report F-score and

One day rodrigo was in the school and he saw a girl and she love her now and they are happy together
So listen Im going to attack ur family with mighty powers.
He's also the best granddaddy ever.
Beatrice Rosen (born 29 November 1985 (Happy birthday)), also known as Batrice Rosen or Batrice Rosenblatt, is a French-born actress. She is best known for her role as Faith in the second season of the TV series "Cuts".

Table 3: Examples of vandalism detected by baseline+PCFG features. Baseline features alone could not detect these vandalism. Notice that several stylistic features present in these sentences are unlikely to appear in normal Wikipedia articles.

All those partaking in the event get absolutely "fritzeld" and certain attendees have even been known to soil themselves
March 10,1876 Alexander Grahm Ball dscovered th telephone when axcidently spilt battery juice on his expeiriment.
English remains the most widely spoken language and New York is the largest city in the English speaking world. Although massive pockets in Queens and Brooklyn have 20% or less people who speak English not so good.

Table 4: Examples of vandalism that evaded both our baseline and baseline+PCFG classifier. Dry wit, for example, relies on context and may receive a good score from the parser trained on regular Wikipedia edits ($C_{regular}$).

AUC rather than accuracy.[1] The baseline system, which includes a wide range of features that are shown to be highly effective in vandalism detection, achieves F-score 52.6%, and AUC 91.6%. The baseline features include all features introduced in Section 3.

Adding language model features to the baseline (denoted as +LM in Table 1) increases the F-score slightly (53.5%), while the AUC score is almost the same (91.7%). Adding PCFG based features to the baseline (denoted as +PCFG) brings the most substantial performance improvement: it increases recall substantially while also improving precision, achieving 57.9% F-score and 92.9% AUC. Combining both PCFG and language model based features (denoted as +LM+PCFG) only results in a slight improvement in AUC. From these results, we draw the following conclusions:

- There are indeed unique language styles in vandalism that can be detected with stylometric analysis.
- Rather unexpectedly, deep syntax oriented features based on PCFG bring a much more substantial improvement than language models that capture only shallow lexico-syntactic patterns.

[1] A naive rule that always chooses the majority class (*"not vandalism"*) will receive zero F-score.

Feature Analysis Table 2 lists the information gain ranking of our features. Notice that several of our PCFG features are in the top ten most informative features. Language model based features were ranked very low in the list, hence we do not include them in the list. This finding will be potentially advantageous to many of the current anti-vandalism tools such as vulgarisms, which rely only on shallow lexico-syntactic patterns.

Examples To provide more insight to the task, Table 3 shows several instances where the addition of the PCFG derived features detected vandalism that the baseline approach could not. Notice that the first example contains a lot of conjunctions that would be hard to characterize using shallow lexico-syntactic features. The second and third examples also show sentence structure that are more informal and vandalism-like. The fourth example is one that is harder to catch. It looks almost like a benign edit, however, what makes it a vandalism is the phrase "(Happy Birthday)" inserted in the middle.

Table 4 shows examples where all of our systems could not detect the vandalism correctly. Notice that examples in Table 4 generally manifest more a formal voice than those in Table 3.

5 Related Work

Wang and McKeown (2010) present the first approach that is linguistically motivated. Their ap-

proach was based on shallow syntactic patterns, while ours explores the use of deep syntactic patterns, and performs a comparative evaluation across different stylometry analysis techniques. It is worthwhile to note that the approach of Wang and McKeown (2010) is not as practical and scalable as ours in that it requires crawling a substantial number (150) of webpages to detect each vandalism edit. From our pilot study based on 1600 edits (50% of which is vandalism), we found that the topic-specific language models built from web search do not produce stronger result than PCFG based features. We do not have a result directly comparable to theirs however, as we could not crawl the necessary webpages required to match the size of corpus.

The standard approach to Wikipedia vandalism detection is to develop a feature based on either the content or metadata and train a classifier to recognize it. A comprehensive overview of what types of features have been employed for this task can be found in Potthast et al. (2010). WikiTrust, a reputation system for Wikipedia authors, focuses on determining the likely quality of a contribution (Adler and de Alfaro, 2007).

6 Future Work and Conclusion

This paper presents a vandalism detection system for Wikipedia that uses stylometric features to aide in classification. We show that deep syntactic patterns based on PCFGs more effectively identify vandalism than shallow lexico-syntactic patterns based on n-grams or contextual language models. PCFGs do not require the laborious process of performing web searches to build context language models. Rather, PCFGs are able to detect differences in language styles between vandalizing edits and normal edits to Wikipedia articles. Employing stylometric features increases the baseline classification rate.

We are currently working to improve this technique through more effective training of our PCFG parser. We look to automate the expansion of the training set of vandalized revisions to include examples from outside of Wikipedia that reflect similar language styles. We also are investigating how we can better utilize the output of our PCFG parsers for classification.

7 Acknowledgments

We express our most sincere gratitude to Dr. Tamara Berg and Dr. Luis Ortiz for their valuable guidance and suggestions in applying Machine Learning and Natural Language Processing techniques to the task of vandalism detection. We also recognize the hard work of Megha Bassi and Thanadit Phumprao for assisting us in building our vandalism detection pipeline that enabled us to perform these experiments.

References

B. Thomas Adler and Luca de Alfaro. 2007. A content-driven reputation system for the wikipedia. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 261–270, New York, NY, USA. ACM.

B. Thomas Adler, Luca de Alfaro, Santiago M. Mola-Velasco, Paolo Rosso, and Andrew G. West. 2011. Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In *CICLing '11: Proceedings of the 12th International Conference on Intelligent Text Processing and Computational Linguistics.*

Wayne Iba Ai and Pat Langley. 1992. Induction of one-level decision trees. In *Proceedings of the Ninth International Conference on Machine Learning*, pages 233–240. Morgan Kaufmann.

Alexa. 2011. Top 500 sites (retrieved April 2011). `http://www.alexa.com/topsites`.

Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Commun. ACM*, 52:119–123, February.

Si-Chi Chin, W. Nick Street, Padmini Srinivasan, and David Eichmann. 2010. Detecting wikipedia vandalism with active learning and statistical language models. In *WICOW '10: Proceedings of the 4rd Workshop on Information Credibility on the Web*.

J. Friedman, T. Hastie, and R. Tibshirani. 2000. Additive Logistic Regression: a Statistical View of Boosting. *The Annals of Statistics*, 38(2).

R. Stuart Geiger and David Ribes. 2010. The work of sustaining order in wikipedia: the banning of a vandal. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, CSCW '10, pages 117–126, New York, NY, USA. ACM.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

Manoj Harpalani, Thanadit Phumprao, Megha Bassi, Michael Hart, and Rob Johnson. 2010. Wiki vandalysis- wikipedia vandalism analysis lab report for pan at clef 2010.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 423–430. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2004. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Polina Panicheva, John Cardiff, and Paolo Rosso. 2010. Personal sense and idiolect: Combining authorship attribution and opinion analysis. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Petra Perner and Sascha Trautzsch. 1998. Multi-interval discretization methods for decision tree learning. In *In: Advances in Pattern Recognition, Joint IAPR International Workshops SSPR 98 and SPR 98*, pages 475–482.

Martin Potthast, Benno Stein, and Robert Gerling. 2008. Automatic vandalism detection in wikipedia. In *ECIR'08: Proceedings of the IR research, 30th European conference on Advances in information retrieval*, pages 663–668, Berlin, Heidelberg. Springer-Verlag.

Martin Potthast, Benno Stein, and Teresa Holfeld. 2010. Overview of the 1st International Competition on Wikipedia Vandalism Detection. In Martin Braschler and Donna Harman, editors, *Notebook Papers of CLEF 2010 LABs and Workshops, 22-23 September, Padua, Italy*, September.

Martin Potthast. 2010. Crowdsourcing a wikipedia vandalism corpus. In *SIGIR'10*, pages 789–790.

Sindhu Raghavan, Adriana Kovashka, and Raymond Mooney. 2010. Authorship attribution using probabilistic context-free grammars. In *Proceedings of the ACL*, pages 38–42, Uppsala, Sweden, July. Association for Computational Linguistics.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *J. Am. Soc. Inf. Sci. Technol.*, 60:538–556, March.

M. Teresa Turell. 2010. The use of textual, grammatical and sociolinguistic evidence in forensic text comparison:. *International Journal of Speech Language and the Law*, 17(2).

William Yang Wang and Kathleen R. McKeown. 2010. "got you!": Automatic vandalism detection in wikipedia with web-based shallow syntactic-semantic modeling. In *23rd International Conference on Computational Linguistics (Coling 2010)*, page 1146?1154.

Andrew G. West, Sampath Kannan, and Insup Lee. 2010. Detecting wikipedia vandalism via spatio-temporal analysis of revision metadata. In *EUROSEC '10: Proceedings of the Third European Workshop on System Security*, pages 22–28, New York, NY, USA. ACM.

Wikipedia. 2010. Daily edit statistics. `http://stats.wikimedia.org/EN/PlotsPngDatabaseEdits.htm`.

Wikipedia. 2011. Wikipedia. `http://www.wikipedia.org`.