Enriching Wikipedia Vandalism Taxonomy via Subclass Discovery

Si-Chi Chin
Interdisciplinary Graduate Program in Informatics (IGPI)
The University of Iowa
Iowa City, U.S.A.
si-chi-chin@uiowa.edu

W. Nick Street
Management Sciences Dept.
IGPI
The University of Iowa
Iowa City, U.S.A.
nick-street@uiowa.edu

Abstract

This paper adopts an unsupervised subclass discovery approach to automatically improve the taxonomy of Wikipedia vandalism. Wikipedia vandalism, defined as malicious editing intended to compromise the integrity of the content of articles, exhibits heterogeneous characteristics, making it hard to detect automatically. The categorization of vandalism provides insights on the detection of vandalism instances. Experimental results demonstrate the potential of using supervised and unsupervised learning to reproduce the manual annotation and enrich the predefined knowledge representation.

1 Introduction

Wikipedia, among the largest collaborative spaces open to the public, is also vulnerable to malicious editing – vandalism. Wikipedia defines vandalism as "any addition, removal, or change of content made in a deliberate attempt to compromise the integrity of Wikipedia[1]." The characteristics of Wikipedia vandalism are heterogeneous. It can a be large-scale editing, such as deleting the entire article or replacing the entire article with irrelevant content. It can be some irrelevant, random, or unintelligible text (e.g. *dfdfefefd #$%&@@#, John Smith loves Jane Doe.*) It can be a small change of facts (e.g. *This is true* $\rightarrow$ *This is not true.*) It can also be an unregulated formatting of text, such as converting all text to the font size of titles. Figure 1 illustrates a taxonomy of Wikipedia actions, highlighting the diverse vandalism instances. The reasons to structure the knowledge of Wikipedia vandalism include:

- sharing common understanding of Wikipedia vandalism,
- making knowledge of Wikipedia vandalism explicit and enabling its reuse,
- providing insights on how vandalism instances are different from legitimate edits, and
- improving the accuracy of Wikipedia vandalism detection.

The detection of Wikipedia vandalism is an emerging research area of the Wikipedia corpus. Prior research emphasized methods to separate the malicious edits from the well-intentioned edits [West *et al.*, 2010; Chin *et al.*, 2010; Smets *et al.*, 2008; Potthast *et al.*, 2008]. Research has also identified common types of vandalism[Vigas *et al.*, 2004; Priedhorsky *et al.*, 2007; Potthast *et al.*, 2008]. However, categorizing vandalism instances relies on laborious manual efforts. The heterogeneous nature of vandalism creates challenges for the annotation process. For example, a "misinformation" vandalism instance can be quite similar to a "nonsense" or a "graffiti" instance [Priedhorsky *et al.*, 2007; Chin *et al.*, 2010]. Current research has yet to establish a standardized or commonly accepted approach to construct the knowledge representation of vandalism instances. In this paper, we introduce an unsupervised learning approach to automatically categorize Wikipedia vandalism. The approach uses statistical features to discover subclasses in both the positive and negative spaces, identifying the partitions that perform the best in multi-class classification. The proposed approach aims to:

- enrich the Wikipedia vandalism taxonomy and knowledge representation automatically,
- improve vandalism detection performance,
- identify potential multi-label instances, and
- identify potential annotation errors.

The paper is structured as follows. In Section 2, we describe the data sets used for our experiments, and detail the implementation of the system. In Section 3 we present our experimental results. In Section 4, we review previous academic research on knowledge representation of Wikipedia vandalism and subclass discovery. In Section 5, we conclude the paper and discuss the opportunities for future work.

2 Experimental Setup

The experiments used the annotated Microsoft vandalism data set provided by Chin et al. [Chin *et al.*, 2010] [2] The dataset has 474 instances with 268 vandalism instances, comprising 21 features extracted from the Statistical Language Model [Clarkson and Rosenfeld, 1997] and Unix *diff* procedure. It also has annotations of 7 types of vandalism : *blanking, large-scale editing, graffiti, misinformation, link spam,*

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

[2] http://code.google.com/p/wikivandalismdata/downloads/list

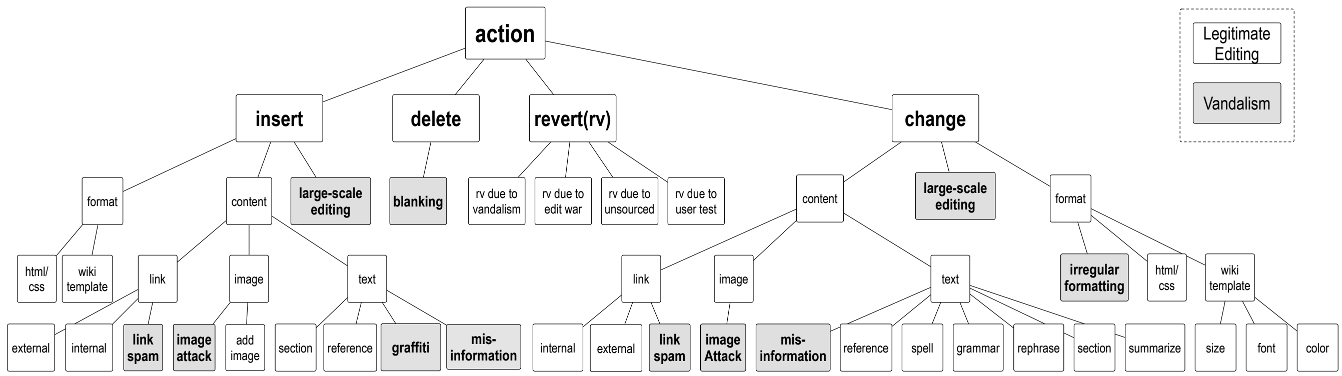


Figure 1: Wikipedia Action Taxonomy. The taxonomy groups Wikipedia editing by the four primary actions (change, insert, delete, and revert) and types of change (format and content), considering also the scale of editing. The shaded boxes are types of Wikipedia vandalism.

irregular formatting, and image attack. The distribution of the 7 types is shown in Figure 3.

Our approach combines unsupervised and supervised learning. Broadly, we use a clustering method to segment both the positive and negative spaces, allowing a better representation for the disjunctive nature of both vandalism and legitimate edits. The cluster memberships are then used as labels in a multi-label classification scheme. Our evaluation, however, is based on the original two labels.

The data was first shuffled into 10 randomized sets. For each shuffle, we clustered the data using k-means clustering. Classification was performed using a support vector machine (SVM) with RBF kernel, using a grid search to find the optimal C and γ parameters. For each shuffle, we used 9/10 of the data as the training set, using the parameters learned from the grid search, to learn a multi-class SVM classifier. The RBF kernel produces a highly nonlinear decision boundary for the disjunctive concept, allowing more accurate results compared to a linear boundary. To evaluate the results, we performed 10-fold cross-validation for each shuffle and averaged the ranked results. The optimal partition was selected based on the average precision (AP) [3] and the Area Under ROC Curve (AUC) metrics. We compared the unsupervised experiment results with the manually annotated results. Figure 2 shows a flowchart of the proposed approach and the design of the experiments. We used Weka [Hall *et al.*, 2009] to implement all experiments.

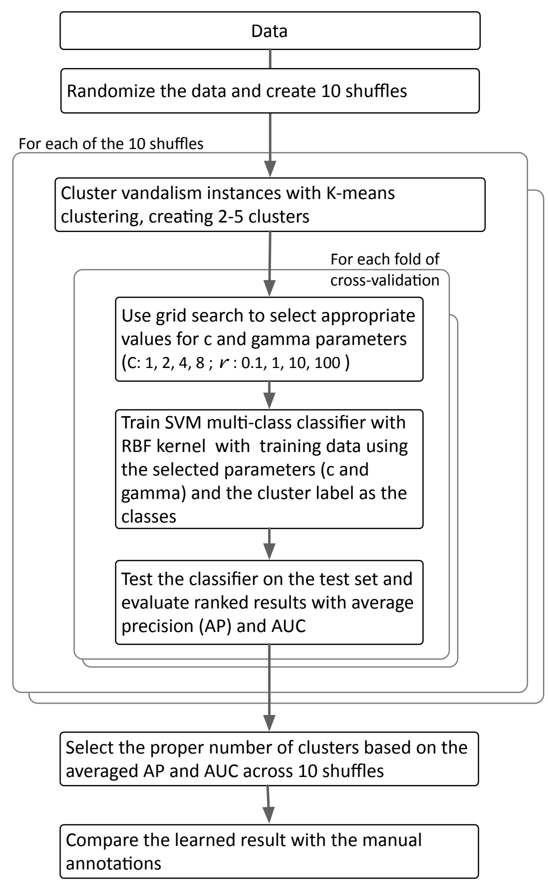


Figure 2: Experiment flowchart

3 Experiment Results

3.1 Unsupervised Clustering vs. Manual Labeling

The experiments used unsupervised clustering to determine the optimal partitions of data that performed the best in the multi-class classification. The clusters were then compared to

[3] We used the following definitions to compute the average precision (AP):

$$AP = \frac{\sum_{r=1}^{N}(P(r) \times \mathrm{rel}(r))}{\textit{number of relevant documents}}$$

$$P(r) = \frac{\textit{relevant retrieved documents of rank r or less}}{r}$$

the manual annotations to explore the opportunity of enriching the predefined knowledge representation of Wikipedia vandalism.

Tables 1 and 2 show the multi-class classification performance for 20 different combinations of positive and negative classes. Both metrics indicate the ideal number of clusters are three for the positive space and four for the negative space. The multi-class classification, compared to the binary classification, increase the AP from 0.425 to 0.443 and the AUC from 0.711 to 0.737. The increases are significant compared to the baseline binary classification.

	P.2	P.3	P.4	P.5
N.1	0.42832	0.43634	0.43874	0.44211
N.2	0.42522	0.43097	0.43720	0.43573
N.3	0.42789	0.43884	0.43538	0.43259
N.4	0.43197	**0.44374**	0.43675	0.43707
N.5	0.42999	0.43878	0.43242	0.43064
Baseline (binary class): **0.42752**				

Table 1: Average Precision (AP) scores of 20 combinations of positive and negative subclasses.

	P.2	P.3	P.4	P.5
N.1	0.71676	0.72800	0.72431	0.72592
N.2	0.71377	0.71648	0.72447	0.72264
N.3	0.71936	0.73366	0.72505	0.72298
N.4	0.72358	**0.73723**	0.72912	0.72640
N.5	0.72434	0.73021	0.72192	0.71693
Baseline (binary class): **0.71144**				

Table 2: Area under curve (AUC) scores of 20 combinations of positive and negative subclasses.

3.2 Enhanced Taxonomy Recommendation

We manually examined the content of vandalism instances in each cluster in order to answer the following questions:

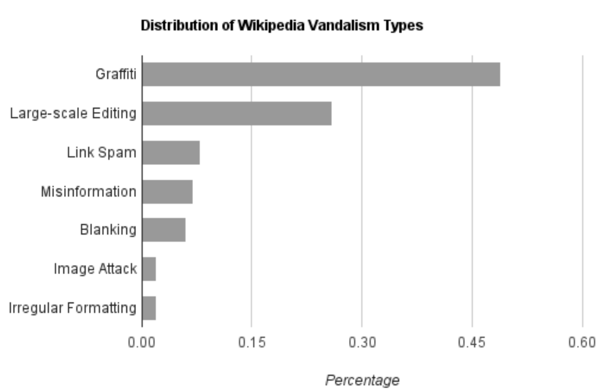


Figure 3: Distribution of Wikipedia Vandalism Types

- How are the instances of large-scale editing and graffiti different from each other in the three clusters?
- Can we identify annotation errors?

Table 3 presents the comparison between the results of clustering and the manual annotation. It is observed that about two-thirds of the graffiti instances are assigned to Cluster 2 with the remaining third assigned to Cluster 3. It is also noted that the large-scale editing instances appeared in all three clusters. The content analysis of the clusters provides insights to enhance the predefined taxonomy, and to discover multi-label instances and annotation errors.

Three Types of Large-scale Editing

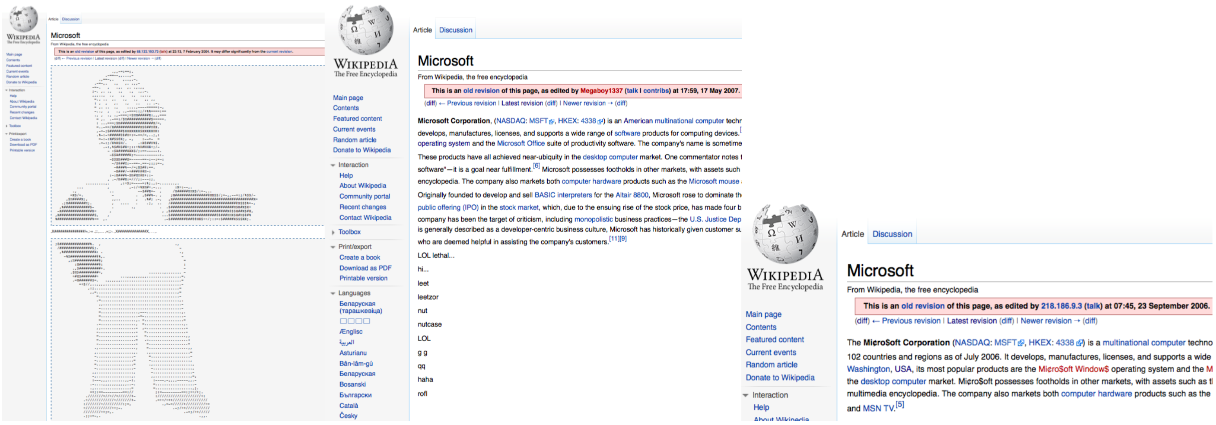
We observed, from Table 3, three different types of large-scale editing. Figure 4 exemplifies three typical instances of large-scale editing from each of the three clusters. The feature space contains three clusters of the large-scale editing instances. We manually examined the data in each cluster to characterize the three types of large-scale editing.

We observed that Cluster 1 contains large insertions of text with diverse vocabulary, usually co-occurring with massive deletion of existing text. For example, we found an ASCII art of the cartoon figure Homer Simpson[4], a complete gibberish text[5], replacing the article with the Apple Computer, Inc article[6], and massive replacement of spelling [7]. Cluster 2 contains the large-scale editing instances that have massive insertion of text with a substantial amount of deletion. [8] [9] Cluster 3 contains instances with numerous spelling changes and named entity replacements, for example, changing "Microsoft" to "Nintendo" ; "Bill Gates" to "George Bush"[10];

[4]http://en.wikipedia.org/w/index.php?title=Microsoft&oldid=2330007
[5]http://en.wikipedia.org/w/index.php?title=Microsoft&oldid=9122754
[6]http://en.wikipedia.org/w/index.php?title=Microsoft&oldid=81420090
[7]http://en.wikipedia.org/w/index.php?title=Microsoft&oldid=9923514
[8]http://en.wikipedia.org/w/index.php?title=Microsoft&oldid=24305432
[9]http://en.wikipedia.org/w/index.php?title=Microsoft&oldid=131585774
[10]http://en.wikipedia.org/w/index.php?title=Microsoft&oldid=62013580

Cluster	Types	Count	Recall
1	**Large-scale Editing** Blanking	28 1	96 %
2	**Graffiti** Misinformation Link Spam Large-scale Editing Blanking Irregular Formatting Image Attack	84 18 15 10 6 2 2	83 %
3	**Graffiti** **Large-scale Editing** Blanking Link Spam Irregular Formatting Image Attack	46 32 9 7 4 4	56 %
Total number of vandalism instances:		268	74 %

Table 3: Cluster analysis of vandalism types



(a) Cluster 1 (b) Cluster 2 (c) Cluster 3

Figure 4: Typical large-scale editing in the three clusters. Edits in Cluster 1 involved large insertion of rich and diverse text. Edits in Cluster 2 involved mass insertion with substantial deletion. Edits in Cluster 3 involved the replacement of named entities and spellings.

"Microsoft" to "Micro$oft."[11]

Two Types of Graffiti: Large vs. Minor Scale

Graffiti is an insertion of unproductive, irrelevant, random, or unintelligible text. We examined the two types of graffiti in Cluster 2 and Cluster 3. We found that graffiti in the Cluster 2 involved insertion of short irrelevant text, such as "LOOK AT ME I CAN FLY!!!![12]" or "I like eggs... [13]". Graffiti in the Cluster 3 involves inserting short unintelligible text, such as "blurrrrrgj[14]," "dihjhkjk, [15]," and "asfasfasf[16]."

Multi-label Instances and Annotation Errors

Although the predefined taxonomy (see Figure 1) considered both the amount of edit (i.e. How much has been changed compared to the previous edits?) and the content characteristics of edits (i.e. What are the edits?), categories that overlap two dimensions are absent in the taxonomy. However, the content analysis indicates numerous instances of copy-and-paste of irrelevant text that has both characteristics of large-scale editing and graffiti [17] [18] [19] [20] [21], as well as massive deletion mixed with graffiti[22] [23].

The results confirm the diverse nature of Wikipedia vandalism, indicating the possibility of improvement for the predefined taxonomy. For example, to include multi-label instances, creating new categories such as "Repeating graffiti (see Figure 5)" to describe the large amount of repeating insertion of irrelevant text, or "Erasure by graffiti" to describe the replacement of majority of content with nonsensical words would enrich the knowledge representation of Wikipedia vandalism.

We searched for the irregular distribution patterns from Table 3 to investigate potential annotation errors. The single blanking instance in the Cluster 1 should actually be a large-scale editing[24]. This finding shows the potential of our approach to amend annotation errors.

4 Related Work

Previous research has identified many common types of vandalism. Viégas et al. [Vigas *et al.*, 2004] identified five

[11] http://en.wikipedia.org/w/index.php?title=Microsoft&oldid=7732342[illegible]

[12] http://en.wikipedia.org/w/index.php?title=Microsoft &oldid=28384195

[13] http://en.wikipedia.org/w/index.php?title=Microsof t&oldid=13233361

[14] http://en.wikipedia.org/w/index.php?title=Microsoft &oldid=86731761

[15] http://en.wikipedia.org/w/index.php?title=Microsoft &oldid=78923750

[16] http://en.wikipedia.org/w/index.php?title=Microsoft &oldid=69519551

[17] http://en.wikipedia.org/w/index.php?title=Microsoft &oldid=45456321

[18] http://en.wikipedia.org/w/index.php?title=Microsoft &oldid=41754476

[19] http://en.wikipedia.org/w/index.php?title=Microsoft &oldid=27056109

[20] http://en.wikipedia.org/w/index.php?title=Microsoft &oldid=12899659

[21] http://en.wikipedia.org/w/index.php?title=Microsoft &oldid=24631945

[22] http://en.wikipedia.org/w/index.php?title=Microsoft &oldid=76785744

[23] http://en.wikipedia.org/w/index.php?title=Microsoft &oldid=63662542

[24] http://en.wikipedia.org/w/index.php?title=Microsoft &oldid=89513491

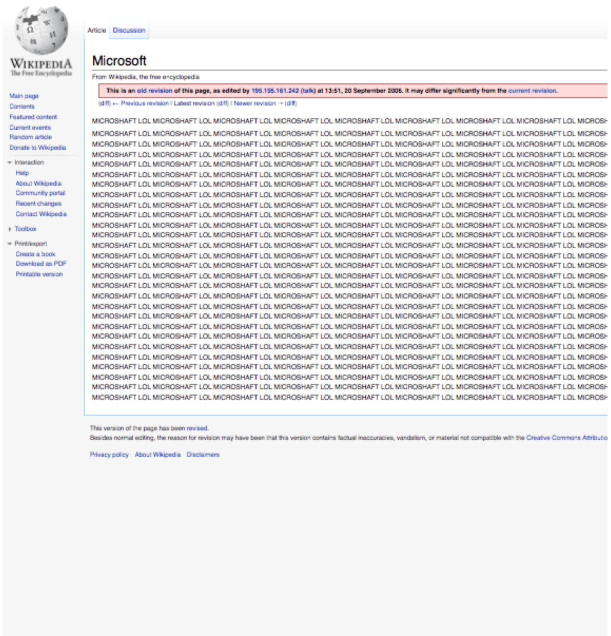


Figure 5: An example of mixed-type graffiti. This instance involves the replacement of entire Microsoft article with repeating nonsensical text.

common types of vandalism: mass deletion, offensive copy, phony copy, phony redirection, and idiosyncratic copy. Priedhorsky et al. [Priedhorsky *et al.*, 2007] categorized Wikipedia damaged edits[25] into seven types: misinformation, mass delete, partial delete, offensive, spam, nonsense, and other. Potthast et al. [Potthast *et al.*, 2008] organized vandalism edits according to the "Edit content" (text, structure, link, and media) and the "Editing category" (insertion, replacement, and deletion). Chin et al. [Chin *et al.*, 2010] constructed a taxonomy of Wikipedia editing actions based on the four primary actions (change, insert, delete, and revert) and types of change (format and content). They identified 7 types of vandalism : *blanking, large-scale editing, graffiti, misinformation, link spam, irregular formatting, and image attack.* The categories proposed in prior works were primarily based on empirical observations of researchers, and can be made more comprehensive or systematically. In our work, we propose using unsupervised clustering and supervised multi-class classification to discover and enrich the knowledge representation of Wikipedia vandalism.

Classification problems involve assigning data to observed categories. In the setting of binary classification, the data has only two classes: positive and negative. However, binary classification becomes difficult in the presence of a heterogeneous positive space. An increasing number of papers have discussed motivations and methods of multi-class classification. [Li and Vogel, 2010a; Lorena *et al.*, 2008; Garca-Pedrajas and Ortiz-Boyer, 2011; Tsoumakas *et al.*, 2010; Zhou *et al.*, 2008]. Subclass classification is subset of multi-class classification, where the multiple class labels belong to a hierarchical structure, and has been shown to enhance classification accuracy. Li and Vogel [Li and Vogel, 2010a; 2010b] utilized sub-class partitions to achieve better performance than the traditional binary classification on the 20 newsgroups dataset. Assent et al. [Assent *et al.*, 2008] incorporated class label information to provide appropriate groupings for classification.

Our work recognizes the heterogeneous nature of Wikipedia Vandalism, discovering clusters that achieved the best performance in the subclass classification. We use the information of discovered subclasses to evaluate and enrich the predefined Wikipedia vandalism categories.

[25] Although damage edits were not referred to as vandalism in their work, they were in fact in line with the definition of Wikipedia vandalism.

5 Conclusion and Future Directions

This paper addresses the problem of detecting diverse Wikipedia vandalism categories, and the problem of recommending appropriate knowledge representation of Wikipedia vandalism instances. We used k-means clustering to map learned categories to a predefined taxonomy, and used supervised classification and content analysis to assist the discovery of novel categories, multi-label instances, and annotation errors.

Wikipedia vandalism detection has previously been regarded as a binary classification problem: ill-intended edits vs. well-intended edits. However, the characteristics of Wikipedia vandalism are in fact heterogeneous. Therefore, our work approached it as a multi-class classification problem, and used unsupervised learning to enhance the manual annotations. Our experimental results showed enhanced performance from the use of multi-class classification method. The results also demonstrated the ability to automate the process of discovering and enriching the Wikipedia vandalism knowledge representations using unsupervised learning.

Future work may include more annotated datasets and comparing the knowledge representation schema between different articles. It is also valuable to investigate how the learned knowledge could be transferred from one articles to the others. Future work may also explore the temporal aspect of the knowledge representation, describing the dynamic evolution of Wikipedia vandalism categories.

References

[Assent *et al.*, 2008] I. Assent, R. Krieger, P. Welter, J. Herbers, and T. Seidl. SubClass: Classification of multidimensional noisy data using subspace clusters. *Advances in Knowledge Discovery and Data Mining*, page 4052, 2008.

[Chin *et al.*, 2010] Si-Chi Chin, W. Nick Street, Padmini Srinivasan, and David Eichmann. Detecting Wikipedia vandalism with active learning and statistical language models. In *Proceedings of the 4th Workshop on Information Credibility*, WICOW '10, pages 3–10, New York, NY, USA, 2010. ACM. ACM ID: 1772942.

[Clarkson and Rosenfeld, 1997] Philip Clarkson and Ronald Rosenfeld. Statistical language modeling using the CMU-Cambridge toolkit. pages 2707—2710, 1997.

[Garca-Pedrajas and Ortiz-Boyer, 2011] Nicols Garca-Pedrajas and Domingo Ortiz-Boyer. An empirical study of binary classifier fusion methods for multiclass classification. *Information Fusion*, 12:111130, April 2011. ACM ID: 1920692.

[Hall *et al.*, 2009] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11:1018, November 2009. ACM ID: 1656278.

[Li and Vogel, 2010a] B. Li and C. Vogel. Improving multiclass text classification with error-correcting output coding and sub-class partitions. *Advances in Artificial Intelligence*, page 415, 2010.

[Li and Vogel, 2010b] B. Li and C. Vogel. Leveraging subclass partition information in binary classification and its application. *Research and Development in Intelligent Systems XXVI*, page 299304, 2010.

[Lorena *et al.*, 2008] Ana Carolina Lorena, Andr C Carvalho, and Jo\ ao M Gama. A review on the combination of binary classifiers in multiclass problems. *Artificial Intelligence Review*, 30:1937, December 2008. ACM ID: 1670491.

[Potthast *et al.*, 2008] Martin Potthast, Benno Stein, and Robert Gerling. Automatic vandalism detection in Wikipedia. In Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and Ryen White, editors, *Advances in Information Retrieval*, volume 4956 of *Lecture Notes in Computer Science*, pages 663–668. Springer Berlin / Heidelberg, 2008. 10.1007/978-3-540-78646-7_75.

[Priedhorsky *et al.*, 2007] Reid Priedhorsky, Jilin Chen, Shyong (Tony) K. Lam, Katherine Panciera, Loren Terveen, and John Riedl. Creating, destroying, and restoring value in Wikipedia. In *Proceedings of the 2007 International ACM Conference on Supporting Group Work*, pages 259–268, Sanibel Island, Florida, USA, 2007. ACM.

[Smets *et al.*, 2008] K. Smets, B. Goethals, and B. Verdonk. Automatic vandalism detection in Wikipedia: Towards a machine learning approach. In *AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy*, page 4348, 2008.

[Tsoumakas *et al.*, 2010] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Mining multi-label data. In Oded Maimon and Lior Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 667–685. Springer US, 2010. 10.1007/978-0-387-09823-4_34.

[Vigas *et al.*, 2004] Fernanda B. Vigas, Martin Wattenberg, and Kushal Dave. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 575–582, Vienna, Austria, 2004. ACM.

[West *et al.*, 2010] Andrew G West, Sampath Kannan, and Insup Lee. Detecting Wikipedia vandalism via spatio-temporal analysis of revision metadata. In *Proceedings of the Third European Workshop on System Security*, EUROSEC '10, page 2228, New York, NY, USA, 2010. ACM. ACM ID: 1752050.

[Zhou *et al.*, 2008] Jie Zhou, Hanchuan Peng, and Ching Y Suen. Data-driven decomposition for multi-class classification. *Pattern Recognition*, 41:6776, January 2008. ACM ID: 1285197.