Identifying controversial articles in Wikipedia: A comparative study

Hoda Sepehri Rad
Department of Computing Science
University of Alberta,Canada
sepehrir@ualberta.ca

Denilson Barbosa
Department of Computing Science
University of Alberta, Canada
denilson@ualberta.ca

ABSTRACT

Wikipedia articles are the result of the collaborative editing of a diverse group of anonymous volunteer editors, who are passionate and knowledgeable about specific topics. One can argue that this plurality of perspectives leads to broader coverage of the topic, thus benefitting the reader. On the other hand, differences among editors on polarizing topics can lead to controversial or questionable content, where facts and arguments are presented and discussed to support a particular point of view. Controversial articles are manually tagged by Wikipedia editors, and span many interesting and popular topics, such as religion, history, and politics, to name a few. Recent works have been proposed on automatically identifying controversy within unmarked articles. However, to date, no systematic comparison of these efforts has been made. This is in part because the various methods are evaluated using different criteria and on different sets of articles by different authors, making it hard for anyone to verify the efficacy and compare all alternatives. We provide a first attempt at bridging this gap. We compare five different methods for modelling and identifying controversy, and discuss some of the unique difficulties and opportunities inherent to the way Wikipedia is produced.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Experimentation

Keywords

Wikipedia, Controversy, Disagreement, Comparison, Monotonicity

1. INTRODUCTION

Wikipedia is arguably the most popular source of knowledge and user-generated content on the Web: at the time of writing, it is among the top-10 most visited sites. Wikipedia's success lies in its distributed and open nature where virtually anyone can become an editor of any article. This openness is both essential for quality control in Wikipedia as well as the source of concern with respect to content *quality*. One aspect of quality is *credibility*: although Wikipedia is generally trustworthy (because of good intentions of most users and because of a dedicated group of administrators that always monitor edits to prevent vandalism and revert inappropriate revisions), it still happens that some revisions are not necessarily trustworthy [1, 10, 12, 23].

Another concern, which is the focus of this paper, is that of *controversy*. Controversy arises as soon as there are sufficiently different and/or contradictory views about a subject, especially when it is hard or even impossible for one to judge where the truth lies. Controversy is unavoidable, as it comes naturally in many topics such as religion, history and politics. Often, opposing views of editors lead to polarization of opinions, resulting in heated disputes rooted at irreconcilable differences in background, belief and perspective. Note that controversy and trust are not synonyms: while it is reasonable to deem a controversial article as untrustworthy, the reverse is not necessarily the case (as there are other reasons that make an article untrustworthy). Unlike trust, controversy arises from the *sequence* of actions and edits in the article, and, thus, does not apply to a single revision of an article.

Compared to the enormous number of studies that have been done on trust, the problem of controversy management has received much less attention from the research community. While the current set of controversial articles (manually tagged by editors) in Wikipedia forms a small fraction of all articles, it should be noted that these articles span a wide range of topics, including many well-known, and very popular subjects (thus attractive to all Wikipedia users). For instance, about 9% of the top visited articles according to statistics obtained in 2010[1] are listed as controversial.

Automatically identifying disputable content within articles can improve the current (manual) process which relies on the editors to tag articles as controversial. It can also be very useful as a tool for analyzing the topics and patterns of collaboration among editors that lead to controversy. Such an analysis can provide more insight about the issues and the positioning of editors that lead to the controversy, which can be helpful for both managing the controversy within the Wikipedia community, and judging the content of articles by readers. For instance, informing the readers of specific issues

WikiSym'12, Aug 27-29, 2012, Linz, Austria.
Copyright 2012 ©ACM 978-1-4503-1605-7/12/08 ...$15.00.

[1] http://stats.grok.se/en/top

about a controversial topic may help them discern whether or not to be cautions and skeptical about certain portions of an article.

Goals. There has been some recent work on automatic detection of controversial content and on quantify their degree of controversy [3, 13, 18, 21, 22]. Most of these methods aim at providing a single controversy score which is then used in classifying or ranking articles. However, to date, no systematic comparison of these efforts has been made. This is in part because the various methods are evaluated using different criteria and on different sets of articles by different authors, making it hard for anyone to verify the efficacy and/or compare different methods. For instance, Brandes et al. [3] studies only 60 random controversial articles, while Kittur et al. [13], and Vuong et al. [22] focused only on articles about religion. Sumi et al. [21] use a simplistic model of controversy, concluding that the complexity of detecting controversy in Wikipedia has been over-estimated and there is no need for designing complex models. However, they neither used a standard evaluation strategy, nor did they compare their results with previously proposed methods such as the work of Kittur et al. [13] and Vuong et al. [22].

Contributions. In this paper, we attempt to close the gaps indicated above. We study and compare different models of controversy under a standard framework and in terms of different metrics. In particular, we show that while methods such as bipolarity and mutual reverts are simple and intuitive, in practice the underlying process of controversy formation in Wikipedia pages is too complex to be captured by these heuristics. Thereby, identifying controversial articles out of a pool of non-controversial articles needs to employ more sophisticated methods such as machine learning tools where controversy is detected by using a combination of factors learned from some annotated examples. On the other hand, while machine learning methods are very powerful at discriminating controversial and non controversial articles, they might be unfavorable depending on how much training samples they require to achieve acceptable performance, and the difficulty in understanding predictions.

We show that these methods can achieve reasonable accuracy even with limited training data. We also argue that in addition to discrimination power, it is desirable for any controversy score to be monotonic. Having the monotonicity property assures that the controversy score assigns higher controversy level to more controversial articles and gives a relatively consistent ranking of articles in terms of their controversy level. Comparing different controversy methods based on this property, some methods including machine learning based methods are shown to not behave like this, which can limit their usability.

Organization. The rest of this paper is organized as follows: in section 2 we give background information about controversy in Wikipedia and the framework used for comparing methods, followed by brief description of each of the methods in section 3. Then, in sections 4, 5, and 6 we report the results of comparison of methods in terms of discrimination, training cost, and monotonicity respectively. In section 7, we discuss challenges and opportunities in modelling controversy in Wikipedia and point out to some possible directions for future. Finally, section 8 concludes the paper.

2. BACKGROUND

2.1 Controversy in Wikipedia

The New Oxford American Dictionary defines *controversy* as "disagreement, typically when prolonged, public, and heated". In the context of Wikipedia, this loosely translates into articles whose edit histories contain one or more "edit-wars" amongst editors. Typically, these articles infringe Wikipedia's Neutral Point of View (NPOV) policy as well. A list of articles with a history of controversy, bias, or NPOV-related issues exists and is maintained[2] by Wikipedia editors manually, who must insert specific templates, referred to as dispute tags [13] to indicate the controversial and disputed state of articles. Examples of these templates are : {{totally-disputed}}, {{disputed-section}}, {{POV}}.

Some authors have looked at bias and the promotion of individual points of views in Wikipedia. Flöck et al. [7] point out several problems such as resistance against new content from "occasional" editors, the difficulty in changing the content in stable and mature articles, and cases with strong feeling of ownership and defensive behaviour of some editors. They argue that such issues affect the diversity and NPOV in Wikipedia. Brandes et al. [5] study some of the factors that lead to editors dropping-out, and show that editors that contribute to controversial articles are more likely to drop-out. One explanation for this phenomenon is the frustration of being involved in long debates, vandalism and full on edit-wars.

The problems mentioned above emphasize the importance of automatic methods to help both editors and readers to understand the controversy process better. Visual analytics are one of the ways towards this goal, where several previous attempts have been made to summarize and visualize revision history and collaboration process of articles. For instance, Kittur et al. [20] use reverts as disagreement relations between editors, and visualize these relations by applying a special layout so that editors with more disagreement appear in longer distances. Similarly, Brandes et al. [4] use short time difference between consecutive revisions as a model of disagreement and propose a visualization technique showing the dominant authors and roles of editors such as "reviser" or "being revised". These tools can help users perceive the confrontation and disagreements between specific groups of editors, but are ineffective in showing the *source* of such conflicts, the degree of controversy, or the points of view of those groups of editors involved in the controversy.

Li et al. [15] verified the source of controversy by testing three hypothesis, namely, a) controversy arises from specific controversial issues covered in an article, b) controversy arises due to the polarizing nature of some topics, and c) the behaviour of a few aggressive editors make some articles to become controversial. They conclude that, in general, controversy is more likely to occur out of the specific issues with the subject of the article.

2.2 Methodology

Besides studying the sources of controversy, it is also interesting to compare the different previous models of controversy against each other on the same corpus of articles. Moreover, it is also useful to see whether intuitive assumptions about controversy hold in Wikipedia. This paper pro-

[2] Wikipedia:List_of_controversial_issues

vides a comparative study of 5 methods for modelling controversy in Wikipedia in a framework of identifying controversial articles. In order to study these methods, we consider the binary classification problem of determining whether or not a specific article is controversial. Our evaluation is based on articles from the list of controversial articles in Wikipedia (the positive positive examples), and articles that neither appear on this list nor have any history of dispute tags or long, debated discussions on their discussion pages as our negative examples (the negative examples).

Some of the methods we evaluate aim at *ranking* articles based on their degree of controversy, (e.g., by predicting the number of dispute tags an article should get). This is a slightly different problem than binary classification: While it is reasonable to assume that an article with many tags is controversial, a low or zero value does not necessarily mean lack of controversy [21]. For instance, many of the articles on the list of controversial articles do not have any dispute tag in their history. Moreover, judging the degree of controversy of articles with tags is problematic because of issues of *disputes over tags* and *over-tagging*[3]. These issues arise when editors disagree on whether or not a tag should be added or removed from an article, and when different, possibly vague and non helpful tags are used for an article.

Two previous studies [13, 22] considered a very limited set of tags (1 and 6 types of dispute tags, respectively), compared to the Wikipedia's currently long and diverse list of *dispute templates*[4]: in at least in 16 of them, the words controversy and dispute are mentioned specifically, and others deal with less explicit forms of controversy. This shows that not only the tag taxonomies and their usage change over time, but also raises concerns to giving equal controversy weights to different tags whose intended meaning are hard to compare. For instance, it is hard to discern whether tags *Cite Check* and *Original Research* rise issues about controversial content, or trustworthiness of the content. Therefore, even though there might be different levels of controversy in different controversial articles, due to lack of reliable ground truth, we study the problem of identifying controversial articles regardless of their degree of controversy.

Classification vs Ranking. By viewing controversy identification as a binary classification task, we need to convert the continuous scores obtained from the output of some of the methods we study. Scores in all of these methods are numeric functions, where higher values indicate more controversy. Mapping continuous outputs to binary outputs is a common problem in the medical domain such as in diagnosing diseases. Suppose $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ are series of examples, where $x \in \chi$ is an instance represented by a continuous score, and $y \in \{-1, 1\}$ represents class labels (i.e. controversial or not). Now, assume f is a decision function that attaches a label y to each instance x as follows:

$$f_a(x) = \begin{cases} 1 & \text{if } x > a \\ -1 & \text{if } x <= a \end{cases}$$

Then, the goal is to find the optimal threshold a that minimizes the misclassification error, which is equal to $P(yf(x) \leq 0)$. There are different classical methods for finding the optimal a such as grid search, ROC curves, and parametric models where a specific distribution such as normal distribution is assumed for the samples [8]. As the ROC curve is a more common method and does not make any assumption about the distribution of samples, we used this method in our work.

In the ROC curve approach, the optimal a is identified by varying the value of a, and calculating the true positive rate, and true negative rates for each value of a. Then, depending on the importance and weights of misclassifications of the two classes, the value of a that maximizes a combination of these two rates is chosen. For our problem, we assigned equal importance to the two classes (controversial and non-controversial), and thereby the optimal a is found for each score-based method at the threshold where the sum of true positive and true negative rates are maximized.

2.3 Metrics

We compare the methods using three criteria:

- Discriminative power: the accuracy of the method in finding controversial articles;
- Cost of training; which approximates the effort from the user before the method can be used; and
- Monotonocity of the method.

3. EXAMINED METHODS

We now discuss the five methods we compare. What is common in all of these methods is that they all rely on simple numeric features extracted from the revision history of the article or article discussion page without analyzing the textual content of the pages.

Table 1 summarizes the main characteristics of the studied methods in terms of the model used for disagreement and controversy.

The following sub-sections give more detailed description of each of these methods.

3.1 Mutual Reverts

Mutual reverts is a single score intended to quantify and rank the degree of controversy of Wikipedia articles [21]. This score relies on revert actions as the direct sign of disagreement and dispute between editors. However, reverts are also common in combating vandalism in non-controversial articles, thus the authors focused only on *mutual reverts*, where two editors have reverted each other's edits at least once. To account for different activity rate of editors, and to filter out less active editors such as vandals, the method considers the minimum number of edited versions of each editor in each pair of mutual revert actions. In this way, disputes with occasional editors such as vandals get less weight than when they occur between more passionate editors. Moreover, the method avoids "personal" conflicts restricted to two specific editors by ignoring the maximum conflict score (of all pairs) within each article. Finally, the total number of distinct editors engaged in mutual reverts is considered as another important factor in heating the debates. The final proposed score is as shown in the following equation:

$$MR^k = E \times \sum_{N_i^k, N_j^k < max} min(N_i^k, N_j^k)$$

[3] Wikipedia:Tagging_pages_for_problems

[4] Wikipedia:Template_messages/Disputes

Table 1: Summary of the main characteristics of the studied methods

Method	Disagreement model	Controversy model
MR	mutual reverts	mutual reverts + editors activities
bipolarity	deletes + reverts	closeness to a bipartite graph
basic	deletes	relative frequency of deletes
structure classifier	learned inferred attitudes	statistics from collaboration networks
meta classifier	-	statistics from article and discussion page

In this equation, MR^k refers to MR score of article k, and N_i^k, and N_j^k are the number of revisions made by two mutually reverting editors i, and j for this article. Also, max is a constant equal to the largest value of the $min(N_i^k, N_j^k)$ across all of these editors to filter out the pair with the maximum conflict score.

This simplistic metric relies on information that is easy to extract: reverts, and the number of edits by each editor. This makes it fast and easy to calculate, compared to other metrics we study in this paper. Also, these simple factors allow the model to work across different Wikipedia languages (for instance compared to discussion page metric which is mostly useful for English Wikipedia where discussion on discussion pages is more common than some other languages). In addition, the authors showed that this simple metric outperforms several different single metrics such as the number of authors or the size of the discussion page in ranking controversial articles.

However, for their evaluation, the authors only considered the precision in the top-30 ranked articles returned by scores of each metric. While it is expected that the top scores arise from controversial articles, precision in mid and low ranges of values were not tested. For instance, the authors reported the percentage of controversial articles for different values of scores. For values below 180, 50% of articles are controversial, while this ratio is 60% for the values under 1000, which shows that in both of these ranges, both controversial and non-controversial articles have the same likelihood. Of course, with increasing the threshold of scores, precision increases, but the recall in the sense of finding examples of controversial articles at the same time decreases. This is why general discrimination ability and considering performance across both classes of controversial and non controversial articles are important for evaluation of such methods.

3.2 Bipolarity

Bipolarity is a single numerical score that is extracted from the “collaboration network” of a Wikipedia article [3]. The collaboration network is a graph where nodes are the editors and edges represent the (positive or negative) relationship between the corresponding editors. The number of words restored from a reverted edit was used as the weight of positive edges, and a combination of the number of reverted edits, and deleted words was considered for negative edge weights.

Having this collaboration network, bipolarity focuses only on negative edges as disagreement relationships between editors. The metric assigns a score between 0 and 1 representing how much the collaboration network of an article is close to a full bipartite graph. The higher the score, the more similar the graph is to a perfect bipartite graph, and the more likely it is to have two opposing camps of editors where most disagreement edges are between the two camps rather than within each.

The authors [3] showed that on the average controversial articles have higher bipolarity scores than featured articles, a group of high quality articles, which is quite consistent with the intuition that one might have about the formation of controversy where bipolar structure of editors is expected. However, as pointed out in [17] the variances of bipolarity scores for these two classes of articles is quite high, which limits the applicability of bipolarity for distinguishing controversial articles.

This limited ability as partly shown in a previous work [18] can be attributed to the approach taken for building collaboration networks. First, bipolarity works only with the negative (disagreement) edges and does not take advantage of positive edges that have been shown to be important in some previous works on signed networks [14]. Second, in inferring the weights of disagreement edges, a simple model based on only deleted words and revert actions have been adopted, which can limit the effectiveness of this method compared to more sophisticated models which infer these edges using a more extensive set of previous behaviour and interaction of editors. In particular, as noted by Brandes et al. both delete and revert actions are seen in featured articles too, where combating with vandalism is a common pattern and these articles also had high bipolarity values. Hence, these actions are at least unable to distinguish between dispute-based disagreements, and vandalism-based disagreements.

3.3 Basic

One of the earliest papers which specifically targeted the problem of controversial articles in Wikipedia is the work of Voung et al. [22], where three different controversy models were proposed. Two of these models consider a controversy score for editors in addition to considering scores for articles, and jointly model these scores in a recursive way. More specifically, they assume that a dispute is more serious when it happens between aggressive and combative editors who have high controversy scores on less controversial articles, or between editors with low controversy scores on articles with high scores. Hence, at each step, the controversy score of an article is updated by the amount of dispute that happened between its editors weighted by their controversy score at that step. Next, the controversy score of editors will be updated based on the updated controversy of edited articles, and this dual updating process continues until scores converge. Dispute between each pair of editors is considered as the number of words that were written by one editor and deleted later by the other editor.

Using the same dispute model, the authors also proposed a simpler, non-reinforcing approach in their paper referred to as *basic* model, where controversy is calculated as follows:

$$C_k = \frac{\sum_{i,j} d_{i,j}^k}{\sum_i o_i^k}$$

where C_k is the controversy score assigned to article k, and $d_{i,j}^k$ and o_i^k are the disagreement values between each pair of editors i and j, and the number of words authored by author i respectively.

Hence, the basic model is just the ratio of deletes to all contributions, where it is expected that controversial articles have higher ratios. The authors showed that the reinforcing-based methods have better performance than this basic model. However, we were unable to successfully apply the reinforcing methods on our dataset. The main reason is that Voung et al. [22] focused on a specific category, where there is a large number of common articles for each pair of editors which makes to have a small number of articles to be processed for each editor in the reinforcing updating procedure.

However, on our dataset, articles were sampled from very different categories, where the chance of finding common articles between target editors (i.e. editors contributed to test articles) is very low. In order to calculate the score of these target editors while recursively calculating the scores of the target test articles, one has to process a very large bipolar graph of articles and editors. For instance, in the first layer which is where we have target editors, we had examples of editors with thousands of edited articles, where expanding these thousand articles at the second layer can add hundreds of thousands of articles and editors.

Even limiting to expanding only a fixed number of articles of each editor instead of expanding all of his articles at each level led to a graph with 85,015 articles, and 241,586 editors for processing testing controversy score of only 480 articles. Unfortunately, our efforts in getting final scores for our test articles even with this limited version failed and we could not get the final scores for our test articles. It should be noted that aside from the excessive computational demand to process and update scores recursively on this big graph, the convergence of scores can be the another reason of our unsuccessful attempt. In particular, these mutual scores do not follow the general template of many HITS-like algorithms as scores do not directly depend on controversy score of the contributed editors, but rather are proportional to an aggregation (such as average) of scores of a *pair* of disputing editors.

Therefore, with not being able to test the reinforcing-based approaches, we focused on studying the basic proposed model which for simplicity is referred to as basic model in the rest of paper.

3.4 Structure classifier

Similar to bipolarity, structure classifier [18] works based on building collaboration networks, but by considering higher-level behavior of editors both in their individual forms, and their pairwise interactions. In particular, a collaboration profile containing these individual and pairwise features is built for each two interacting editors and is classified as positive or negative denoting the general agreement or disagreement relation of them. For this classification, votes in admin elections were used as training labels.

Once the collaboration profiles of each two interacting pairs was built and classified, a collaboration network consisting of these classified profiles as signed edges is built for each article similar to bipolarity method. However, compared to bipolarity, in this method both positive and negative edges are considered for modelling controversy. In addition, collaboration networks are not represented by a single metric, but rather by extracting the following groups of features from each network.

- basic features such as number of nodes, number of positive edges, etc.
- degree distribution features such as the percentage of nodes having an in-degree of higher than 90% of maximum in-degree, and similarly for out-degree
- triad features where triads are subgraphs of size 3 with 8 different types

Since, networks are represented by a feature vector, the final controversy model is based on a classification approach where the feature vector representation of collaboration networks are learned to be controversial or not.

Hence, the structure classifier works in two training phase and by using two classifiers: one for classifying profiles which is used for assigning the sign of edges of collaboration networks, and another for classifying networks into controversial or not labels.

3.5 Meta classifier

The meta classifier proposed by Kittur et al. [13] is another classification approach for identifying controversial articles which relies on extracting a set of objective statistics from the revision history of an article or from its discussion page. The authors proposed 30 different features such as the number of revisions of an article, the number of unique editors, the number of out-link, and in-links, etc. and found the following 7 features as the most important features:

- The most important features found for meta classifier (# means number of).
- #revisions of the discussion page
- #minor edits of the discussion page
- #unique editors of the discussion page
- #revisions of the article
- #unique editors of the article
- #revisions of the discussion page by anonymous editors
- #revisions of the article by anonymous editors

While none of the other methods considers the discussion page associated to an article in modelling controversy, we can see that more than half of the most important features in the meta classifier are related to statistics of such pages. Kittur et al. [13] also emphasized the importance of discussion pages in their paper by showing that there has been a trend of less direct edits on articles, and instead more edits and discussion on discussion pages. However, this trend only was studied on English Wikipedia and as suggested by some other works, discussion pages are less active in some other languages like Romanian and Persian [21].

As mentioned before, Kittur et al. [13] considered the number of controversial tags to evaluate their method. In order to study methods from the binary classification perspective and to avoid problems discussed in section 2 about relying on controversial tags, we use the same features, but instead of learning a regressor, we learned a binary classifier using controversial or non-controversial labels assigned to articles. Hence, instead of using the SMOLogistic method which is a regression model used in [13], we used the SMO classifier (we tried other classifiers too, but this one gave the best results for the considered features).

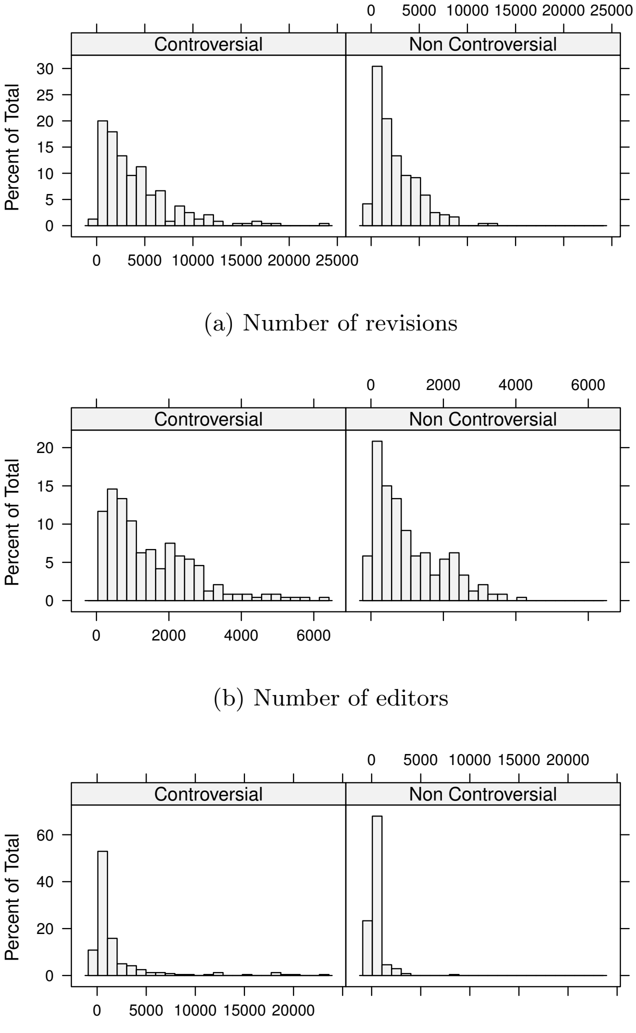
Finally, it should be noted that a different meta classifier was later proposed in [17] using some of the meta features from Kittur et al. and some new features, where all features were extracted from only the revision history of the article itself. We tested this classifier too, and it achieved a slightly higher accuracy than the original meta classifier. However, as the work of Kittur et al. [13] is representative of one of the well-known early studies on controversy in Wikipedia, and is it the only method that considers discussion pages prominently, we focus on that meta classifier.

4. DISCRIMINATIVE POWER

In this section, we compare different models in terms of their effectiveness in distinguishing controversial from non-controversial articles, which we refer to as the *discriminative* power of the methods. The results we report were obtained on the same dataset of 240 controversial and 240 non-controversial articles used in a previous work [18]. In this dataset, the controversial articles were randomly selected from all 15 categories in the list of controversial articles, and the 240 non-controversial articles were randomly selected from high-quality (i.e.featured) and mid-quality articles by assuring that they never appeared in the list of controversial articles, or have any controversial tag in the article or corresponding discussion pages.

Baselines. We use the following baselines as means of comparison: (1) the number of unique editors contributing to each article (#editors); (2) the number of revisions of each article (#revisions); and (3) the number of revisions of discussion page associated with each article (#talk-revisions). Intuitively, one might expect that a large group of editors, a high number of revisions, or a long history of discussion should be indicative of controversy. However, even though controversial pages usually have large number of these factors, none of these factors alone are sufficient for telling apart controversial articles from other articles. For instance, long history of discussions is also common in featured articles, even though the goal is usually different from debating, and discussions are more centered around activities for improving the article coverage and style of writing.

Figure 1 shows the distribution of controversial and non-controversial articles in our test set in terms of each these factors. As can be seen, there are some examples of controversial articles with high values of the mentioned factors, but there are plenty of other examples that lay within the same range as non-controversial articles. Comparing the three factors, #talk-revisions is more discriminative than the other two, but still the samples of both classes are widely spread out and do not form a distinct boundary. It should be noted that these three baselines are among the top-7 ranked features in the meta classifier [13], where a combination of these factors along with other features are used in a classification approach to overcome the limited discrimination of each individual baseline.



(a) Number of revisions

(b) Number of editors

(c) Number of revisions in talk page

Figure 1: Article distribution by baseline.

Results. Table 2 compares performance of the five studied methods along with the baselines based on the overall accuracy, and also on a per-class accuracy. Per-class accuracy corresponds to the number of instances of each class that are correctly classified. In the binary classification literature, these are usually referred to as *sensitivity* and *specificity* for the positive and negative classes respectively. In some applications such as predicting a disease based on test results, the risk of misclassification of one class is assumed to be higher than the other and more weight might be given to improve the accuracy within that class. However, for our application, it is unclear whether there is any difference in the cost of misclassification between the classes, and thereby a method which generates a higher accuracy while providing a good balance in the accuracy of both classes is preferred.

Table 2: Comparison of overall and per-class accuracy

method	accuracy	cont-acc	ncnt-acc
MR	0.67	0.55	0.76
basic	0.60	0.83	0.36
bipolarity	0.56	0.57	0.49
meta classifier	0.75	0.86	0.63
structure classifier	0.84	0.86	0.83
#talk-versions	0.64	0.42	0.85
#article-versions	0.57	0.46	0.68
#editors	0.56	0.45	0.67

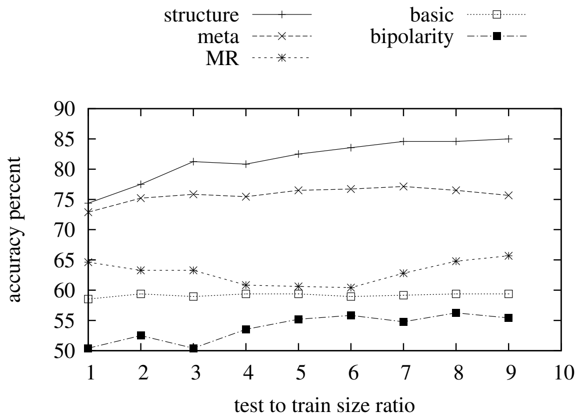


Figure 2: Effect of training size on accuracy

The results of comparing methods first shows that the two classifier-based methods have the best overall accuracy among the studied methods, where specially there is a large gap between the performance of the structure classifier and all the other methods. This highlights the importance of combining several different indicators and employing machine learning-based methods.

Also, note that #talk-versions is a baseline that has an overall accuracy higher than some of the methods, such as bipolarity and basic. More interestingly, this baseline has the highest accuracy for the non-controversial class. On the other hand, methods such as basic and the meta classifier have a much higher accuracy on the controversial class. In practice, we want a classifier to achieve good results in both classes. In this term, the structure classifier is a very successful method that not only has the highest accuracy overall, but its per-class accuracy for both classes is among the highest values across all methods.

5. TRAINING COST

This section studies the effect of the amount of training data on the accuracy of the methods. The costs of collecting training data and training a model are usually very high as they typically involve human efforts. Therefore, it is natural to seek trade-offs between accuracy and amount of training data.

It should be noted that the cost of applying a model is not limited only to the cost of providing training samples and can be extended to the cost of complexity and the availability of required resources to extract features and statistics related to that model. For instance, extracting a feature like the number of unique editors is much easier than features such as the number of articles linking to an article (as in meta classifier), or using a Wikipedia specific resource such as election data to infer the attitudes (as in structure classifier). However, due to difficulty in objectively comparing these different factors, we can only analyze the cost from the perspective of number of training examples.

Unlike classification methods, the score-based methods do not require a training phase as they just assign a score to each sample. However, as explained in section 2, in order to maximize the accuracy of these methods, an optimal cutoff value when scores are mapped to decision labels is needed. The optimal threshold found on a set of labeled data and used to label a set of unseen samples can differ from one sample data to another which affects the accuracy of predicted labels on test data. Hence, we studied the effect of training sample size for these methods too.

Also note that structure classifier has two phases where in the first phase a vote classifier is trained to learn the relationship between collaboration profiles and votes, and in the second phase this learned relationship is used to build collaboration networks and classifies them using structural features into controversial or not. Hence, in principle, the training data of this method cannot be compared with other methods that their training only depends on the article samples. However, as the vote classifier needs to be run only once and is independent from the second training phase, we considered the first training phase as an additional cost factor not studied in the current experiment as this experiment only focuses on the trend of the results with respect to the number of training articles.

Results. Figure 2 shows the trend of accuracy of different methods when trained using data with increasing size, and tested on a fixed dataset. More specifically, we first partitioned the original dataset containing 480 articles into 90% and 10% training and test data respectively. Then, using that fixed 10% of test data, we tracked the accuracy of each model by training using only $n * 10\%$ ($\{n = 1, 2..10\}$) of the original training data. Finally, similar to cross-validation experiments, to reduce variability of results, we did 10 rounds of partitioning the original data , where in each round, we chose a different partition as the test data, and considered the rest as the full training data. Therefore, the accuracy result of each training size was obtained by averaging the results over the 10 rounds. Also, in generating a sample training at each training size and generating a test set at each round we kept the same ratio of controversial and non controversial articles.

As can be seen, using more training data, in general has a positive effect on the accuracy of all methods. However, relative benefits from using more training data differ across methods. For instance, the accuracy of the structure classifier increases by more than 15%, while the basic method has the least increase which is almost independent of the size of training data. In general as expected, the classifier-based approaches are more sensitive to the amount of training data, while score-based methods show less than 5% difference for their results with the change of training size.

What is more interesting is that even when using just 10% of the available training data, both structure and meta classifiers achieved a very reasonable accuracy of about 75%, which is 10% higher than the best score-based methods. Moreover, the relative performance of all methods remained the same, regardless of the amount of training data. This serves as a strong argument in favor of classification-based methods from the discrimination aspect.

6. MONOTONICITY

We say that controversy score $C(\cdot)$ fulfills the *monotonicity criterion* if it assigns less or equal score to an article p if some parts of the article were removed from it. That is, if p' is obtained by excluding some parts of the content of p, then, $C(p') < C(p)$ for C to satisfy the monotonicity criterion. The intuition behind monotonicity is that removing some parts of an article cannot increase the **absolute** global controversy level of that article, as doing so can only remove some of the sources of dispute.

We posit that monotonicity and discrimination power are complementary qualities in a controversy detection method: a method satisfying both conditions would be useful in applications beyond the simple binary task of identifying controversial articles. For instance, a monotonic controversy score could be used to find *particularly controversial sections* within an article. A manual analysis of some controversial articles, and the results of source of controversy hypothesis done by Li et al. [15] show that there are in fact only some specific parts in controversial articles, which are the focus of debates and disputes. For instance, sections about *abortion and breast cancer* and *questioning the authorship* of some of the works assigned to Shakespeare brought up most of the conflicts in the *abortion* and *Shakespeare* articles. Being able to specify specific controversial issues beyond classifying articles to controversial or not can give a better insight about the controversial articles to readers and the editor community itself to manage and resolve conflicts.

One approach towards extracting issues in controversial articles can be to look for a set of units in each article that by removing them, the article will not be considered as controversial based on a reliable controversial method. These set of units should be chosen in an optimal way so that a unit will be only removed if it contributes to the total controversy score of the article, so that non-controversial units will not be chosen. In this way, a search over all possible combinations of units for finding the optimal set of units using heuristic search algorithms will benefit if the controversy score that is used to lead the search direction decreases monotonically. Regardless of the design and the complexity of such a search algorithm, which is beyond the scope of this paper, monotonicity seems to be essential for any non brute-force approach.

In addition monotonicity test can be a general indication of whether a controversy score can rank different articles according to the level of controversy and conflict that they experienced. It is expected that any controversy ranking method to generate monotonic scores at both article, and unit level.

Results. We test the methods for monotonicity using 50 controversial articles randomly chosen from the set of 240 controversial articles, using entire sections as text units. For each article, we rank the units according to the number of edits they comprise. We successively remove units, in decreasing ranking, and measure the controversy score obtained from each method after each removal, recalculating all features as necessary.

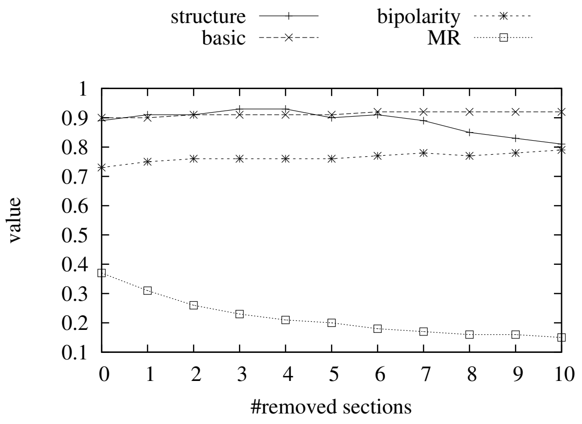


Figure 3: Monotonicity test for the studied methods

For the meta classifier, which relies on the discussion pages, updating features extracted from the discussion page is challenging as it requires mapping edits on the article to comments in the discussion page. However, the discussion page might not form the same topic granularity of the sections of the article or might be described by different wordings. Hence, we exclude the meta classifier from the monotonicity analysis in the present paper. Future work can address this issue by finding a mapping between these two sets of features, or defining the monotonicity condition considering the discussion and the article pages together.

Also, for having a controversy score for the structure classifier, we rely on the membership probability of the controversial class, assigned by this classifier to each test instance (i.e. article). For getting these probabilities, we used the Weka machine learning package.

Figure 3 shows the changes of absolute controversy scores of different methods as more and more sections are removed. As we can see, MR is the only method that strictly decreases by removal of sections and satisfies the monotonicity condition. Other scores-based methods, namely bipolarity and basic do not satisfy these conditions. This can be attributed to the normalized, and relative nature of these methods. For instance, in the basic method, as sections are removed, the sum of disagreement values ($d_{i,j}$) decreases, but at the same time, the sum of authorship values (o_i) decreases too. Depending, on the size of the drop in sum of disagreement values, and the size of the drop in the sum of authorship values, the ratio of these two quantities can increase, decrease, or stay constant.

Also, for the structure classifier, we see that in general it is not monotonic and initially has a small increasing trend, and then decreases by higher rate after removal of the fifth section. This behaviour can be attributed to the possibility that as we remove sections and update the features of articles to get the modified instances, the instances get farther and farther from the space of the instances the classifier has seen in its training phase. In fact, the modified instances might be not only different from their original controversial training instances, but also not similar to non-controversial training instances. This way, the probabilities obtained from the classifier become less and less reliable with the removal of sections, and the assumption of modelling class membership and controversy score based on them will not be correct.

Hence, application of classifier-based methods for ranking problems such as the ones we discussed faces the challenge of being able to define a reliable class membership and controversy score for all test instances.

7. DISCUSSION

Wikipedia is one of the well-known examples of social media which has been studied based on different aspects in recent years. Analyzing controversy in this medium faces a set of challenges which are more or less specific to this domain. Specially, identifying controversial cases based on revision history of articles requires detecting arguments and opinions that can be expressed in a more implicit way. An editor might express his disagreement by simply deleting some content without replacing it with an alternative, or providing any reasoning for his actions. Also, arguments, debates, and biases can be implicit in the differences of two snippets that are very similar. For instance, a snippet of "Most scholars believe that ..." was changed to "Some scholars believe that..." to change the tone and support of an opinion in a Wikipedia article. These implicit, subtle disagreement cases are in contrast to the different vocabulary usage of different opinion camps that were found in forums or news [6, 9, 19].

Also, making sense out of a high volume and continuous, fast pace of changes which are mixed with vandalism is another challenge that makes Wikipedia different from domains where a limited, specific sets of posts, or news articles should be processed. On the other hand, by allowing public access to log of all activities of editors, Wikipedia (and similar wiki systems) provides a valuable source of knowledge that is hardly seen in other domains.

Besides the score-based vs. classification-based categorization of controversy models discussed through out different experiments in previous sections, depending on the considered aspects and resources used, these models can be further categorized into the following four groups:

- **Meta-driven:** the methods in this group rely on extracting a set of numeric, simple statistics from the revision history of the article or/and its discussion page where these statistics can be combined into a score or a set of feature vectors to be learned in a machine learning framework as the meta classifier, or in a rule-based system;

- **User-driven:** in this category controversy is modeled based on editors interactions and their positive or negative collaborations where the structure classifier, and bipolarity are examples of models building a network of editors based on a notion of agreement/disagreement between editors and extracting a score or set of features respectively. Another example of methods in this category is the mutual reinforcing model of Vuong et al. [22] where the interaction of editors are modeled by the number of words they deleted from each other and the controversy of an article is calculated based on an aggregation of the controversy scores of each two pairs of interacting editors (i.e. calculated in a recursive way).

- **Content-driven:** The third category of methods are methods modelling controversy by analyzing the content of revisions, comments, or the discussion pages. The content analysis can ignore the semantic by applying simple content analysis such as tracking authorship and deleted words in the revision history of the article like in the basic method. Alternatively, the content analysis can depend on semantic of the text by applying Natural Language Processing techniques such as textual entailment of changed versions, or discourse analysis of the discussion pages which with some recent attempts on annotation of discussion pages [2, 16] seems more practical than before.

- **Pattern-driven:** the basis of methods in this group is analyzing patterns of edits over a history of revisions. The MR method that looks at mutual reverts in the revision history as sign of edit wars is an example of these methods. In a more advanced level, these edit patterns are modeled by network motifs in a recent work [11], where the network motifs are defined by considering the network of editors and articles over each three consecutive versions. The frequency of different network motif types (more than 39'000 different types) over the entire revision history of articles is extracted as feature vectors and different edit patterns are learned for controversial and non-controversial articles. Other edit patterns considering more abstract and general types, variable pattern length, and possibly unsupervised extraction of patterns can be studied in future.

Modeling controversy can also be improved by taking advantage of multiple categories and combining different sources. For instance, combining a meta classifier-based method with structure classifier as a user-driven method was shown to be superior to both of these individual methods in a previous work [18]. As another example, a user-driven model can be built by inferring the type of relations between editors based on a content-driven approach such as analyzing the comments or discussions of the corresponding editors in discussion pages.

In addition to these possible improvements, the research on analysis of controversy in Wikipedia can be extended by giving more insight about controversy beyond simply identifying controversial articles. For instance, as briefly mentioned in section 6 ranking controversy level of different text units of an article and identifying the most contested issues can be an interesting direction for this kind of work. The opposing views and positioning of editors towards each of these issues can give more insight about controversial topics.

8. CONCLUSION

In this paper, we studied five different controversy models in Wikipedia in terms of their discriminative power, the cost of learning the models, and the monotonicity condition. The results show that in practice the underlying principles of interaction of editors and the formation of controversy are too sophisticated to be captured by single heuristics and a combination of different factors need to be considered.

In terms of monotonicity, we found out that most methods including classifier-based methods did not satisfy this property. Non-monotone behavior of a controversy model can limit its usefulness when in addition to discrimination, ranking of different articles, or different parts within the same article is needed.

Hence, considering the four different categories of modelling controversy we discussed, the future work can focus on designing models that satisfy both discrimination, and monotonicity condition and apply these models in ranking problems. Other directions we identify are using text summarization techniques to provide concise descriptions of the opinions expressed by opposing groups of editors. Such information, together with an assessment of the severity of each controversial topic in an article could greatly enhance the experience of regular Wikipedia users.

Acknowledgements

This work was supported in part by grants from NSERC.

9. REFERENCES

[1] B. T. Adler, K. Chatterjee, L. de Alfaro, M. Faella, I. Pye, and V. Raman. Assigning trust to Wikipedia content. In *proceedings of the 4th international Symposium on Wikis*, pages 1–12. ACM, 2008.

[2] E. M. Bender, J. T. Morgan, M. Oxley, M. Zachry, B. Hutchinson, A. Marin, B. Zhang, and M. Ostendorf. Annotating social acts: authority claims and alignment moves in Wikipedia talk pages. In *proceedings of workshop on Languages in Social Media*, pages 48–57. ACL, 2011.

[3] U. Brandes, P. Kenis, J. Lerner, and D. van Raaij. Network analysis of collaboration structure in Wikipedia. In *proceedings of the 18th international conference on World Wide Web*, pages 731–740. ACM, 2009.

[4] U. Brandes and J. Lerner. Visual analysis of controversy in user-generated encyclopedias. *Information Visualization*, 7(1):34–48, 2008.

[5] U. Brandes and J. Lerner. Is editing more rewarding than discussion? a statistical framework to estimate causes of dropout from Wikipedia. In *proceedings of workshop on Motivation and Incentives*. ACM, 2009.

[6] Y. Choi, Y. Jung, and S.-H. Myaeng. Identifying controversial issues and their sub-topics in news articles. In *proceedings of Pacific Asia workshop on Intelligence and Security Informatics*, pages 140–153. Springer, 2010.

[7] F. Flöck, D. Vrandečić, and E. Simperl. Towards a diversity-minded Wikipedia. In *proceedings of the 3rd international conference on Web Science*. ACM, 2011.

[8] K. Hajian-Tilaki, J. Hanley, L. Joseph, and J.-P. Collet. A comparison of parametric and non-parametric approaches to roc analysis of quantitative diagnostic tests. *Medical Decision Making*, 17(1):94–102, 1997.

[9] A. Hassan, V. Qazvinian, and D. Radev. What's with the attitude?: identifying sentences with attitude in online discussions. In *proceedings of international conference on Empirical Methods in Natural Language Processing*, pages 1245–1255. ACL, 2010.

[10] M. Hu, E.-P. Lim, A. Sun, H. W. Lauw, and B.-Q. Vuong. Measuring article quality in Wikipedia: models and evaluation. In *proceedings of the 16th international conference on Information and Knowledge Management*, pages 243–252. ACM, 2007.

[11] D. Jurgens and T.-C. Lu. Temporal motifs reveal the dynamics of editor interactions in Wikipedia. In *proceedings of the 6th international conference on Weblogs and Social Media*. AAAI, 2012.

[12] A. Kittur, B. Suh, and E. H. Chi. Can you ever trust a wiki?: impacting perceived trustworthiness in Wikipedia. In *proceedings of the 13th international conference on Computer Supported Cooperative Work*, pages 477–480. ACM, 2008.

[13] A. Kittur, B. Suh, B. A. Pendleton, and E. H. Chi. He says, she says: conflict and coordination in Wikipedia. In *proceedings of the 25th international conference on Computer/ Human Interaction*, pages 453–462. ACM, 2007.

[14] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *proceedings of the 20th international conference on World Wide Web*, pages 641–650. ACM, 2010.

[15] C. Li, A. Datta, and A. Sun. Mining latent relations in peer-production environments: a case study with Wikipedia article similarity and controversy. *Social Network Analysis and Mining*, pages 1–14, 2011.

[16] J. Schneider, A. Passant, and J. Breslin. A qualitative and quantitative analysis of how Wikipedia talk pages are used. In *proceedings of the 2nd international conference on WebScience*, pages 1–7. ACM, 2010.

[17] H. Sepehri Rad and D. Barbosa. Towards identifying arguments in Wikipedia pages. In *proceedings of 20th international conference on World Wide Web: Posters*, pages 117–118. ACM, 2011.

[18] H. Sepehri Rad, A. Makazhanov, D. Rafiei, and D. Barbosa. Leveraging editor collaboration patterns in Wikipedia. In *proceedings of the 23rd international conference on Hypertext and Social Media*, pages 13–22. ACM, 2012.

[19] S. Somasundaran and J. Wiebe. Recognizing stances in online debates. In *proceedings of the 47th annual meeting of the ACL*, pages 226–234. ACL, 2009.

[20] B. Suh, E. H. Chi, B. A. Pendleton, and A. Kittur. Us vs. them: Understanding social dynamics in Wikipedia with revert graph visualizations. In *proceedings of the 2nd international conference on Visual Analytics Science and Technology*, pages 163–170. IEEE, 2007.

[21] R. S. Sumi, T. Yasseri, A. Rung, A. Kornai, and J. Kertész. Edit wars in Wikipedia. In *proceedings of the 3rd international conference on Social Computing*, pages 724–727. IEEE, 2011.

[22] B.-Q. Vuong, E.-P. Lim, A. Sun, M.-T. Le, and H. W. Lauw. On ranking controversies in Wikipedia: models and evaluation. In *proceedings of the 1st international conference on Web Search and Data Mining*, pages 171–182. ACM, 2008.

[23] H. Zeng, M. A. Alhossaini, L. Ding, R. Fikes, and D. L. McGuinness. Computing trust from revision history. In *proceedings of the 4th international conference on Privacy, Security and Trust*, pages 1–8. IEEE, 2006.