Detecting Controversy on the Web

Shiri Dori-Hacohen and James Allan
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts Amherst
Amherst, Massachusetts
{shiri, allan}@cs.umass.edu

ABSTRACT

A useful feature to facilitate critical literacy would alert users when they are reading a controversial web page. This requires solving a binary classification problem: does a given web page discuss a controversial topic? We explore the feasibility of solving the problem by treating it as supervised k-nearest-neighbor classification. Our approach (1) maps a webpage to a set of neighboring Wikipedia articles which were labeled on a controversiality metric; (2) coalesces those labels into an estimate of the webpage's controversiality; and finally (3) converts the estimate to a binary value using a threshold. We demonstrate the applicability of our approach by validating it on a set of webpages drawn from seed queries. We show *absolute* gains of 22% in $F_{0.5}$ on our test set over a sentiment-based approach, highlighting that detecting controversy is more complex than simply detecting opinions.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Storage and Retrieval —*Query formulation, Information filtering*

Keywords

controversy detection, sentiment analysis, critical literacy

1. INTRODUCTION

Publishing material about controversial issues is of paramount importance to a functioning democracy, as it allows disagreements to be aired in public. However, when searching for discussion of a controversial issue it is all too easy to cherry-pick from the results. For example, those against gun rights will surely find material supporting this position (the tragedy of school shootings), whereas those for gun rights will find other evidence (the Second Amendment in the U.S.). At the same time, people searching for Issels Treatment will find a convincing web site describing this "comprehensive immunotherapy for cancer"; yet it is listed as a "dubious treatment" by Quackwatch [11] and the American Cancer Society considers it unproven and maybe harmful [2]. An unsuspecting reader who has not heard of the controversy is likely to be misled or uninformed. Even careful readers suffer from a "filter bubble" [9] wherein automatic and social systems guide readers toward what they expect, feeding into confirmation bias rather than encouraging them to seek the multiple perspectives available on a subject.

We are interested in techniques that encourage and facilitate healthy debates, allowing users to critically approach these issues. One way to do so is to alert users when their search results represent a perspective on a controversial issue; for example, imagine a warning presented at the top of a web page: "This webpage represents one of several perspectives on a controversial topic." To do so, we need to answer a non-trivial question: "Is this topic controversial?"

Note that our goal differs from "diversifying" search results, wherein – perhaps – each of the perspectives might be presented in a ranked list. Instead, we aim to identify whether a single page *in isolation* discuses a topic with wide ranging perspectives. This study is an early investigation into whether that challenge can be solved.

We approach this as an estimation problem: determining the level of controversy in a topic, while thresholding it for binary classification. We utilize a supervised k-nearest-neighbor classifier on web pages that uses labeled estimates of controversy in Wikipedia articles to determine the likelihood that a web page is controversial itself. Essentially, a page similar to controversial pages is likely controversial itself. Our choice of Wikipedia articles as labeled neighbors is motivated both by topical coverage, as well the possibility of using unsupervised labels of controversy from prior work.

We use a collection of 377 web pages that were manually judged as controversial or not. Our approach yields $F_{0.5}$ of 64.8% and accuracy of 72.9% for our test set. Hypothesizing that controverial material is often highly opinionated, we compare our results to a sentiment analysis classifier; we outperform it consistently on all metrics but recall.

2. RELATED WORK

To the best of our knowledge, this problem has not been formulated as such before, though several special cases have been explored by previous researchers.

Controversy Detection in Wikipedia. Early work on detecting controversy focused on Wikipedia, where structured data and revision history provide powerful scaffolding, simplifying detection [7]. Wikipedia pages manually tagged as controversial are a valuable resource, but using the manual tags alone can be problematic due to inconsistency and sparseness of tagging [10, 14]; thus, identifying controver-

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

ACM 978-1-4503-2263-8/13/10 ...$15.00.
http://dx.doi.org/10.1145/2505515.2507877.

Table 1: Data set size and annotations

Webpages			
Set	Seeds	Pages	Controversial
Training	Wikipedia	248	74 (29.8%)
Testing	Wikipedia	129	49 (38.0%)

Wikipedia articles (labeled data)		
Articles	Annotated	Controversial
8,755	1,761	282 (16.0%)

sial Wikipedia articles that have not been manually tagged adds value. Recent work reexamined the variety of machine learning and handcrafted approaches previously published, and offers different criteria: one suggested metric, "M", neutralizes vandalism, which was cited in prior work as a confounding issue in Wikipedia [14]; another paper leveraged collaboration networks between individual editors to identify controversy, with significant improvements reported [12].

Controversy on the web. Our goal is to widen the scope of controversy detection to the entire web. Work on controversy outside Wikipedia has made progress on targeted domains, e.g. Twitter [10] and news [3, 5]; they largely considered politics and politicians. In our case, we would like to approach all controversies, whether political, medical, or religious. The closest work to ours creates a collection [3]; we detect controversy in isolation and in ad-hoc situations.

Recent work includes diversifying search results for controversial queries [6], with less focus on detection that a query is controversial in the first place. Reliance on sources such as Debatepedia[1] [3, 6] presupposes that the debate has been covered. Yet debate websites focus on political issues; as of this writing Debatepedia has no entry discussing Homeopathy. We consider the problem of detecting controversy to have potential utility as a precursor step in diversifying controversial queries, though that is not our main focus.

Sentiment analysis. One approach that can apply to controversy is sentiment analysis, used to detect words that indicate high polarity and opinion [5, 10]. However, unlike sentiment, "controversies are much more complex and opinions are often expressed in subtle forms, which makes determining pro/con polarities much more difficult than [...] prior work on opinion mining" [3, p. 523]. We compare our approach to a baseline sentiment analysis system [1] and show that its performance is lower for this task, yet it has high recall and bears further investigation (see Section 4.3).

3. EXPERIMENTAL SETUP AND DATA SET

To investigate the feasibility of our approach, we construct a suitable data set. We hypothesize that we can detect controversy indirectly by using the controversiality of Wikipedia articles that are similar to the starting webpage. Thus, our data set also includes judgments on the controversiality of Wikipedia articles.

Our data set, described in Table 1, was created as follows. We selected 41 seed articles from Wikipedia. The articles were chosen based on their implied level of controversy, with some clearly controversial ("Abortion") and others clearly not controversial ("Mary Poppins"). We used only the Wikipedia article's title as a query to the blekko search engine[2]. From up to top 100 results returned for queries, we selected only webpages that also appeared in ClueWeb09 category B[3] to allow reproducibility. We also omitted Wikipedia articles, pages that could not be displayed properly, and pages that had no nearest neighbors among the Wikipedia articles (see below and Section 4.2), leaving 377 web pages over the 41 seed topics.

We split this collection into training and testing sets based on the seeds – since our pages were not chosen independently. We wanted approximately a 60-40 split, so we divided our seeds randomly into 30% whose "related" webpages were labeled as all training, 20% as all testing, and 50% of the seeds whose webpages were split, as one group, at a 60-40 ratio between the training and testing collections. The final distribution of the collections differed slightly due to our selection method, as shown in Table 1: the training set had a lower proportion of controversial pages than the testing set (29.8% vs. 38.0%).

We created an annotation tool to capture the controversy level of these pages. We ask how controversial is the topic discussed by the webpage, and the options were: "1 - clearly controversial", "2 - possibly controversial", "3 - possibly non-controversial", or "4 - clearly non-controversial". By design, 344 of the 377 pages were annotated by more than one annotator for 851 total judgments. Table 2 summarizes the agreement among the annotators. 65.1% of the pages had complete agreement, accounting for 64.7% of the judgments. Another 17.4% had a majority (2 of 3) vote, with 17.4% of the pages tied among two annotators.

Our approach also relies on labeled data from Wikipedia. We used a variation of the annotation tool to judge the controversiality of Wikipedia articles. For each of the 377 pages we found its nearest Wikipedia articles using queries to blekko (as described in Section 4.2), for a total of 8755 unique Wikipedia articles. We annotated as many top-ranking Wikipedia articles as we could, resulting in 1761 Wikipedia articles judged by our annotators, as shown in Table 1. Of these, 331 were annotated by more than one annotator, and they agreed on 81.6% of the Wikipedia pages.

Whenever a webpage or Wikipedia article was annotated more than once, we took the average value of all the judgments (in the range [1..4]) as its controversy score, which we use in our approach and evaluation. To convert into a binary value, any score below a threshold of 2.5 (the midpoint of our 4-point range) is considered controversial.

4. EVALUATION

We evaluate our approach as a binary classifier model, where a page is classified as *controversial* or *not controversial*. For this approach, the set marked as controversial by the system can be compared to the truth set described earlier. We calculate precision, recall, F_1, $F_{0.5}$, and accuracy.

4.1 Baseline runs

As a new problem, no obvious baseline algorithm exists. However, since controversy can arguably be described as the presence of strong opposing opinions, a natural baseline is a sentiment analysis classifier. For our baseline, we took a modified version of a state-of-the-art sentiment classifier, a logistic regression model on sentiment features [1]. The only modification is the division into classes, since we are most interested in the presence of sentiment, not its direction; thus,

[1]http://dbp.idebate.org/

[2]http://blekko.com

[3]http://lemurproject.org/clueweb09/

we train a binary classifier in which positive, negative, and mixed sentiments are considered one class ("sentiment") and neutral sentiment the other ("neutral"). The sentiment class is taken as controversial; the neutral, as noncontroversial.

Table 2: Inter-annotator agreement. ***Results are shown separately for 2 and 3 annotators that rated the same page.***

All (2 or 3)	Pages		Judgments	
Total	344		851	
Agreement	224	(65.1%)	551	(64.7%)
Disagreement (all)	120	(34.9%)	300	(35.3%)
2 Annotators	**Pages**		**Judgments**	
Total	181		362	
Agreement	121	(66.9%)	242	(66.9%)
Disagreement (Tie)	60	(33.1%)	120	(33.1%)
3 Annotators	**Pages**		**Judgments**	
Total	163		489	
Agreement	103	(63.2%)	309	(63.2%)
2-1 Disagreement	60	(36.8%)	180	(36.8%)

As two additional baselines, we generated random values as estimates of controversy. One random function assigns equal probability to controversial and noncontroversial pages ("Random$_{50}$"), and another assigns controversy based on the incidence in the training set, i.e., 29.8% ("Random$_{29.8}$"). For each random approach, we averaged the scores of 3 runs. Finally, we also use a dominant class baseline, which judges every webpage as noncontroversial, and thus has zero precision and recall but non-zero accuracy.

4.2 Nearest Neighbor approach

Our approach maps a webpage to a set of Wikipedia articles, and uses the controversiality of those articles to predict whether the page at hand is controversial or not. We use a supervised approach that uses our annotators' judgments of Wikipedia articles to create an estimator of controversy for the webpage, which we then convert to a binary value.

Our starting point is a webpage, from which we automatically generate a query by selecting the top ten non-stopword terms from that page. As mentioned above, we use these terms to query Wikipedia (via blekko), and eliminate any user or talk pages. As mentioned in Section 3, we have labels of controversy on 20% of these articles (with preference towards articles ranking higher in the retrieval). We use our annotators' judgments of Wikipedia articles whenever they are available. We aggregate the score over k neighbors of the webpage to receive a final controversy score. As mentioned in Section 3, we convert the score to binary using a threshold of 2.5.

We vary 4 different parameters in our runs:

1. Stop set: We used two stop sets, the 418 INQUERY stop set [4] or a short, 35 term set (*"Full"* vs. *"Light"* stop).

2. k: we control for the number of neighboring Wikipedia articles used in the calculation. We used [1..20], and one run with no limit (all available matching articles are used).

3. Handling non labeled data: We use two alternatives to "fill in the blanks" when labeled data was not available: One guesses a score of 2.5 for absent neighbor labels, and the other guesses a score of 2.5 for web pages where no neighbors were labeled. However, both versions achieved similar scores, and were identical for all the runs presented; we thus omit this parameter in the remainder of our paper.

4. Aggregate function: we use one of three methods to aggregate the k neighbors' scores: *Max*, *Average*, and *Exponential Average* - an average weighted by $\frac{1}{2^{rank}}$.

All in all, we had 252 parameterized runs (2 stopping options × 2 methods × 3 aggregate functions × 21 limit values). In order to choose the best parameters, we ran a parameter sweep on the training set.

4.3 Results

Table 3 shows the results of the parameter sweep, with runs optimized for P, R, F_1, Accuracy and $F_{0.5}$ on the training set. The upper half of the table presents scores on the training data (the runs used to select the parameters among the 252 possibilities) and the parameters that achieved those scores. The lower part of the table shows the evaluation measures for those same parameters on the test set. All 4 baselines are presented for each of the sets.

5. DISCUSSION

Looking at the results in Table 3, we note that scores on the test set are consistently higher than baseline runs for all metrics but recall. We observe that the runs optimizing for precision and recall on the training set – top two rows of Table 3 – remained stable in the test set. The run optimizing for precision in training also outperformed other runs for Accuracy and $F_{0.5}$ in the test set. In all cases, we present $F_{0.5}$ in addition to F_1; we prefer higher precision over recall.

The results for the test set are in line with the training results, indicating that our method is successful in detecting webpages with controversial topics. The best run overall (Light, k=4, average) achieves 21.9% and 21% absolute gain in $F_{0.5}$ over the sentiment and random baselines respectively. Accuracy is 10.9% higher than the best baseline.

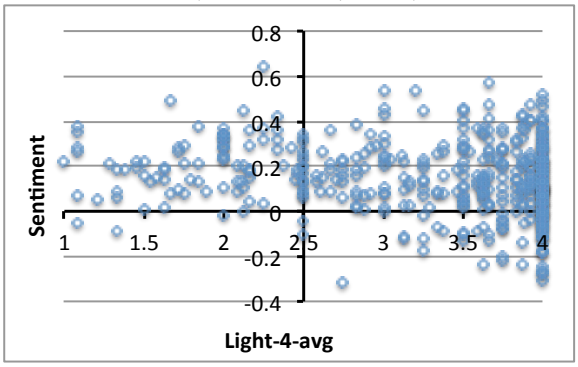
In all runs, we used a threshold value of 2.5 on both our estimator and annotations to create binary judgments. Using the threshold on the estimator was also validated by running a Precision-Recall curve on the training set.

Table 3: Results for the best methods, optimizing for P, R, F_1, Accuracy and $F_{0.5}$ on the training set; presented on both sets. ***Bold cells in the training set represent the metric (column) optimized by the specific parameter run (row); bold cells in the testing set represent the best result of all the runs presented. Baselines are bold whenever they are greater or equal to the best system result.***

Set	Parameters			Results				
	Stop	k	Agg	P	R	F_1	Acc.	$F_{0.5}$
Training Set	Light	4	avg	**0.791**	0.459	0.581	0.802	0.691
	Light	20/no	max	0.555	**0.824**	0.663	0.750	0.593
	Full	9	max	0.703	0.703	**0.703**	0.823	0.703
	Full	6/7	max	0.708	0.689	0.699	**0.823**	**0.704**
	Full	8	max	0.708	0.689	0.699	**0.823**	**0.704**
	Sentiment			0.346	**0.959**	0.509	0.448	0.397
	Random$_{50}$			0.322	0.518	0.397	0.526	0.348
	Random$_{29.8}$			0.141	0.151	0.145	0.631	0.143
	Dominant			0	0	0	0.702	0
Testing Set	Light	4	avg	**0.694**	0.510	0.588	**0.729**	**0.648**
	Light	20/no	max	0.512	**0.837**	0.636	0.636	0.556
	Full	9	max	0.585	0.633	0.608	0.690	0.594
	Full	6/7	max	0.589	0.673	0.629	0.698	0.604
	Full	8	max	0.596	0.694	**0.642**	0.705	0.614
	Sentiment			0.379	**0.898**	0.533	0.403	0.429
	Random$_{50}$			0.420	0.531	0.468	0.545	0.438
	Random$_{29.8}$			0.229	0.190	0.207	0.606	0.220
	Dominant			0	0	0	0.620	0

While the Full stopping runs achieved higher F's and accuracy on the training set, they were less stable across the folds. In our analysis we found that we had a lower propor-

Figure 1: Scatter Plot between Sentiment and system scores. ***Presented for Light-4-avg run on all sets. Sentiment ranges from -1 (neutral) to 1 (sentiment); system scores range from 1 (controversial) to 4 (noncontroversial).***



tion of labeled data among the Wikipedia articles from Full stopping (25.8% compared to 42.6% for Light stopping); this resulted in a higher reliance on estimating unlabeled data. Among Light stopping runs alone, the Light-4-avg runs optimized accuracy and $F_{0.5}$ in training, which is consistent with that run achieving the best test results. An additional run not presented, which optimized F_1 in training among the Light stopping runs (k=3, max aggregate function), was more stable with respect to F_1 than the Full Stopping runs.

The sentiment classifier has high recall but low precision; not every sentiment implies controversy. While the sentiment classifier achieves P, $F_{0.5}$ and accuracy scores that are sometimes comparable to random, its extremely high recall nonetheless leads to F_1 scores that are above random. This consistently high recall suggests it may be valuable as a classification feature; we found that sentiment scores were not correlated with controversy scores, suggesting that combining the two may yield improvement, as shown in Fig. 1.

In addition to the results presented here, we ran our approach on another set of webpages, using a small set of queries from the TREC Blog track [8] as additional seeds. The results for this set (not presented here) were lower for both our approach and all baselines; however, the set was too small to draw any significant conclusions from the results. Additionally, we also tried two unsupervised approaches: the first, based on the presence of dispute tags manually added to the article [7, 12], and the second using the "M" score as defined by Yasseri et al. [14]. However, in our experiments these unsupervised approaches were not successful, in some cases performing worse than random (results not presented here). We found that the scores in both these methods did not line up with our Wikipedia annotations; we have several hypotheses for these results that we hope to explore soon, but which are beyond the scope of this short paper.

6. CONCLUSIONS AND FUTURE WORK

We showed that a supervised nearest-neighbor approach can be used to detect whether a webpage discusses a controversial topic. We map the page to Wikipedia articles and use annotated data to estimate their controversy levels, then using those scores to produce a controversy score for the original webpage. Our results demonstrate that related Wikipedia pages can be used to detect controversy, with our method achieving considerable improvements compared to both a sentiment-based classifier, and random and dominant class baselines.

The benefit of using Wikipedia articles as neighbors lies in both coverage, as well as the potential for unsupervised approaches to be substituted for the supervised estimates of Wikipedia article controversiality. We plan to look into unsupervised approaches for controversy detection such as dispute tags and the "M" metric [14], and analyze why our attempts failed; we will also investigate additional approaches such as a "meta" classifier [7] and collaboration networks [12].

Our current approach to find neighbors can also be improved by using a state-of-the-art research search engine, or using other methods of matching webpages to Wikipedia [13]. The sentiment classifier, with its excellent recall, may be useful as a feature. We would like to address topics that are not covered in detail in Wikipedia, but are nonetheless controversial. On a larger scope, we would also like to go beyond detecting controversy at the topic level, to detect stances and alignment of a specific page to that topic.

7. ACKNOWLEDGMENTS

We thank E. Aktolga, H. Sepehri Rad, T. Yasseri and E. Yom-Tov for fruitful discussions and resources. Thanks to CIIR lab members and the anonymous reviewers for their comments. This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF grant #IIS-1217281. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

8. REFERENCES

[1] E. Aktolga and J. Allan. Sentiment diversification with different biases. In *Proc. SIGIR*, pages 593–602, 2013.

[2] American Cancer Society. Metabolic therapy, March 2012. Retrieved from http://goo.gl/5bWoX.

[3] R. Awadallah, M. Ramanath, and G. Weikum. Harmony and dissonance: organizing the people's voices on political controversies. In *Proc. WSDM*, pages 523–532, 2012.

[4] J. P. Callan, W. B. Croft, and S. M. Harding. The INQUERY retrieval system. In *Database and Expert Systems Applications*, pages 78–83. Springer Vienna, 1992.

[5] Y. Choi, Y. Jung, and S.-H. Myaeng. Identifying controversial issues and their sub-topics in news articles. In *Proc. PAISI*, pages 140–153, 2010.

[6] M. Kacimi and J. Gamper. MOUNA: mining opinions to unveil neglected arguments. In *Proc. CIKM*, pages 2722–2724, 2012.

[7] A. Kittur, B. Suh, B. A. Pendleton, and E. H. Chi. He says, she says: conflict and coordination in Wikipedia. In *Proc. CHI*, pages 453–462, 2007.

[8] Ounis, M. de Rijke, C. Macdonald, G. Mishne, and I. Soboroff. Overview of the TREC-2006 Blog Track. In *Proc. TREC*, 2006.

[9] E. Pariser. *The Filter Bubble: What the Internet is hiding from you.* Penguin Press HC, 2011.

[10] A.-M. Popescu and M. Pennacchiotti. Detecting controversial events from twitter. In *Proc. CIKM*, pages 1873–1876, 2010.

[11] Quackwatch. A special message for cancer patients seeking "alternative" treatments, August 2010. Retrieved from http://goo.gl/ZezZm.

[12] H. Sepehri Rad and D. Barbosa. Identifying controversial articles in Wikipedia: A comparative study. In *Proc. WikiSym*, 2012.

[13] L. Ratinov, D. Roth, D. Downey, and M. Anderson. Local and global algorithms for disambiguation to wikipedia. In *Proc. HLT*, 2011.

[14] T. Yasseri, R. Sumi, A. Rung, A. Kornai, and J. Kertész. Dynamics of conflicts in Wikipedia. *PLoS ONE*, 7(6):e38869, 2012.