Controversy Detection in Wikipedia Using Collective Classification

Shiri Dori-Hacohen*, David Jensen†, James Allan*

*Center for Intelligent Information Retrieval, †Knowledge Discovery Laboratory
College of Information and Computer Sciences
University of Massachusetts Amherst
{shiri,jensen,allan}@cs.umass.edu

ABSTRACT

Concerns over personalization in IR have sparked an interest in detection and analysis of controversial topics. Accurate detection would enable many beneficial applications, such as alerting search users to controversy. Wikipedia's broad coverage and rich metadata offer a valuable resource for this problem. We hypothesize that intensities of controversy among related pages are not independent. Thus, we propose a stacked model which exploits the dependencies among related pages. Our approach improves classification of controversial web pages when compared to a model that examines each page in isolation, demonstrating that controversial topics exhibit homophily. Using notions of similarity to construct a subnetwork for collective classification, rather than using the default network present in the relational data, leads to improved classification with wider applications for semi-structured datasets, with the effects most pronounced when a small set of neighbors is used.

1. INTRODUCTION

Critical literacy, civic discourse and trustworthy information are not immediate results of effective information retrieval. Controversies proliferate online, but the "filter bubble" effect encourages confirmation bias by offering users the answers they want to hear [11]. Exposure to diverse opinions can potentially improve civic discourse, but these benefits will only be available to users who can detect controversial topics. Automated tools performing such detection can support users in their browsing and search experience [6].

Prior work on controversy detection focused on Wikipedia (cf. [9, 12]), analyzing each page in isolation or studying its editors. We hypothesize that controversies occur in neighborhoods of related topics. Thus, advanced ML techniques that take relational data into account, such as *collective* and *stacked* inference [7, 10], can improve controversy detection by exploiting the dependencies among related pages. In a departure from most work on collective inference, we also hypothesize that a definition of "relatedness" that incorporates textual or topical similarity will hold more predictive power than pre-existing relationships such as hyperlinks. If so, using a constructed network based on similarity will outperform collective models based on explicit relations. Our research questions are: What is the relative performance of intrinsic versus collective classification for detecting controversy in Wikipedia? And, what types of page relationships most improve the classification? We hypothesize that collective models will substantially outperform intrinsic models, and particularly when related pages are defined in terms of their textual or topical similarity to the classified page.

SIGIR '16, July 17 - 21, 2016, Pisa, Italy

ISBN 978-1-4503-4069-4/16/07…$15.00

DOI: http://dx.doi.org/10.1145/2911451.2914745

2. RELATED WORK

The related work falls broadly under three themes: the need for controversy detection, methods for controversy detection, and collective and stacked inference.

The Need for Controversy Detection. Increasing personalization reduces exposure to diverse opinions, which is a serious risk for the tenets of deliberative democracy. Search engines and social media use personalization to tailor results to the users' opinions, creating a "Filter Bubble" [11] which can further exacerbate confirmation bias. It is increasingly evident that digesting material about controversies is a challenging task for end users. These concerns have sparked controversy analysis and detection, a research area of growing interest (for a survey of prior work, challenges and important implications, see Dori-Hacohen et al.[6]). Accurately and automatically distinguishing between controversial and noncontroversial topics is one such challenge which is currently within technical reach, yet far from a solved problem. Our paper focuses on automatically detecting controversial topics in Wikipedia, a task proposed by Kittur et al. [9], which can also serve as a crucial step in other algorithms (cf. [4]).

Methods for Controversy Detection in Wikipedia. Of the relatively sparse prior work on automatically detecting controversy, most focuses on Wikipedia, since its rich user-generated content base offers a wealth of semi-structured data (for a survey and comparative study, see

Table 1: Data set size and annotations (Wikipedia Articles)

Set	Articles	Controversial
DHA [5]	1926	293 (15.2%)
SRMRB [12]	480	240 (50%)

Algorithm 1 Cross-validation stacked training procedure

for fold $i = 1..k$, $Set_i = A \setminus \text{fold}_i$ **do**
  Train IM_i, an intrinsic model on Set_i
  Select $subneighbors(Set_i) \subseteq neighbors(Set_i)$
  Apply IM_i on $subneighbors(Set_i)$
  Aggregate predictions of $subneighbors(Set_i)$ to create an extended feature set, Set'_i
  Train SM_i, a stacked collective model on Set'_i
end for

Algorithm 2 Cross-validation stacked inference procedure

for fold $i = 1..k$ **do**
  Select $subneighbors(fold_i) \subseteq neighbors(fold_i)$
  Apply IM_i (trained above) on $subneighbors(fold_i)$
  Aggregate predictions of $subneighbors(fold_i)$ to create an extended feature set, $fold'_i$
  Apply SM_i (trained above) on $fold'_i$
end for

[12]). Most of this work has used an approach that classifies each page in isolation [9, 14]. In contrast, this paper examines networks of pages that are topically related, and argues that controversy detection can be improved by considering a page in the context of its neighbors. While some recent work has alluded to the possibility that controversies occur in neighborhoods of related topics [4] or demonstrated such clusters anecdotally [8], this potential connection has yet to be tested or used to improve controversy detection.

Web-page Classification and General Collective Classification Approaches. Collective and relational inference are ML techniques that can be applied to relational data, which have been successful on many complex problems such as hyperlink categorization [3], by exploiting homophily between related objects [7]. Stacked models are a type of collective classification that avoids the need for computationally intensive inference procedures, and is particularly useful in situations where there is a lack of extensive ground truth data for the neighborhood of a page [10]. In stacked models, an *intrinsic classifier*, relying only on the features of the data instance being evaluated, is trained first, and then applied to generate predictions for the neighbors of every instance in the set. These predictions are then aggregated into an extended dataset and used as features of the instance. Finally, a *stacked model* is trained by using this extended dataset, as in regular collective inference. In other words, the collective inference classifier is "stacked" over the intrinsic classifier (see e.g. Algorithms 1 and 2 below). Instead of using known truth labels of neighbors, a stacked model uses the outputs of an intrinsic classifier. Stacked models have been demonstrated to be effective at collective classification due to a reduction in bias [7].

When stacked models are used in semistructured datasets, they are usually applied in a relational manner: relatedness is defined directly in the structured data. In several domains, however, a relational link between two objects does not imply a strong connection between them. Inspired by the needs of our task, we propose *explicitly constructing* a subnetwork of relationships for the purpose of improving stacked classification. Using features of the semi-structured dataset, such as relation directionality and object similarity, we construct a more useful notion of relationship; we thus depart from most stacked classification approaches that assume that the dataset contains a fixed relational schema (cf. [7, 10]). Our work is distinct from Probabilistic Similarity Logic [2], which reasons about similarity for inference purposes; we propose to construct an induced subgraph of relationships based directly on similarity measures.

Table 2: Intrinsic and Stacked features

Type	Description
Intrinsic	# of Revisions; # of Minor Revisions; # of Editors; # of Anonymous Editors; # of Anonymous Revisions; % of Anonymous Editors; % of Anonymous Revisions; Max Edits Per Editor; Avg Edits Per Editor; Std Dev of Edits Per Editor
Stacked	Proportion of neighbors above X% probability, where X = {10%, 30%, 50%, 70%, 90%}. This represents a discretized version of the probability distribution of controversy among the neighbors; Max Controversy probability among neighbors; Avg Controversy probability among neighbors

3. APPROACH

We will classify Wikipedia pages as controversial or not, using a combination of intrinsic features of a page, as well as predictions of controversy from pages related to it. There are two novel parts to our approach (described below): first, we construct a subnetwork of relations based on similarity, and then proceed to use a stacked model on top of this constructed network. The training procedure for the intrinsic model is the standard fashion. Following Kou and Cohen [10], our stacked training procedure creates neighbor predictions in a cross-validated manner with 10 folds. The main difference from their approach is the use of a subset of the neighbors, rather than all neighbors. The training procedure is applied to the i-th fold, as seen in Algorithm 1. At inference time, the stacked model pipeline is applied to the i-th fold in an analogous manner, as seen in Algorithm 2.

Constructing a Subnetwork. We examine the neighborhood of each Wikipedia page, for stacked classification and to evaluate whether homophily exists for controversial topics. The effectiveness of collective inference relies on homophily between related instances. Presumably, if a page is controversial, then the pages related to it are likely to be controversial. The controversy level of related pages, therefore, can be used as a feature to the collective model. However, links in Wikipedia are noisy, and not necessarily the best indication of relatedness. We expect stacked classification to be more useful when applied specifically to more relevant links. We thus do not consider every hyperlink to be an equally valid neighbor, but instead apply a similarity function to generate a relative ranking among all neighbors. Additionally, we argue that links pointing into, and out of, an article, should be viewed as separate types of relationships. Incoming links consist of a zipfian-like distribution which grows on a logarithmic scale, while outgoing links exhibit a more linear relationship. Specifically, we construct a subnetwork by applying a TF-IDF-based pairwise cosine similarity function on the text of the page, and then selecting the top-scoring neighbors (taken as two separate lists, for in-links and out-links) as most "related" to the center page.

Creating a Stacked Model. To evaluate our hypotheses, we create intrinsic and collective models of controversy.

Table 3: Compared Systems

Name	Description
Stacked-Ranked-k	Proposed stacked inference system with a similarity-based subnetwork
Intrinsic	A classifier using only intrinsic features
Stacked-All	A stacked inference system, as above, but which uses all Wikipedia neighbors
Stacked-Random-k	A stacked inference system which uses k randomly selected neighbors
Neighbors-Only-k	A classifier based only on the neighbor predictions (as in a regular stacked model), without using the intrinsic features of the center page
Prior work	See Sepehri Rad & Barbosa [12] for details

We compare an *intrinsic classifier* that classifies each page independently, and a **collective inference classifier** that assumes dependence between controversy values of related pages.

4. DATASET AND EXPERIMENTAL SETUP

We would like to examine the following hypotheses: (1) using a subset of chosen neighbors, based on a similarity ranking, represents an improvement upon using all neighbors; (2) using this subset also represents an improvement upon using the same amount of random neighbors. We will describe the datasets used, the model features and setup, and finally the alternative systems we created in order to examine our hypotheses.

4.1 Data Sets

We use two datasets for this work, as described in Table 1, which were created by two independent groups. The first dataset is the publicly available[1] Wikipedia Web Controversy dataset (denoted DHA [5]). The second is a collection provided on request (denoted SRMRB [12]). The incidence of controversy is different in the two sets (about 15% in DHA and exactly 50% in SRMRB). While it is quite challenging to estimate the precise incidence of controversy in the wild, we believe that an unbalanced setting is more realistic - in general, noncontroversial topics far outnumber controversial topics. In order to partially mitigate the challenges of training on an imbalanced set (DHA), we applied weights to all the instances in the training folds, such that the sum of weights of all controversial pages was equal to the sum of weights for the noncontroversial pages.

4.2 Model Features and Setup

For both the intrinsic and the stacked models, we use the Random Forest classifier provided by Weka, set to use 100 trees, and the default behavior for all other settings. For training and inference, we used 10-fold cross-validation, as described in Section 3.

Similarity for Subnetwork Construction. In order to generate the collective model, we observed all Wikipedia pages linking into, and out of, the center page. We ranked all these pages by pairwise, TF-IDF based cosine similarity (ignoring stop words), then chose the top k in-links and the top k out-links of the central page. We considered several alternatives for thresholding the similarity. In the experiments described below, we simply pick the top k ranked neighbors for incoming links, as well as the top k for outgoing links, where k is either 10 or 300.

Features. The features of both of the intrinsic and stacked models are displayed in Table 2. *Intrinsic Features* follow prior work that used metadata features of the Wikipedia pages [9, 12]. All intrinsic features are extracted from the May 2014 Wikipedia dump [2]. A subset of the features were extracted using JWPL[3]. We use the intrinsic model to generate predictions (probabilities of controversy) for each neighbor in the subnetwork described above. Collective inference requires that the relevant features of pages be aggregated in order to use them: we use the aggregate functions in Table 2, applied separately to in-links and out-links. In total, 14 *Stacked Features* were added (7 aggregates each, which were applied to the top k in-links and out-links separately).

Table 4: Results for compared models with $k = [10, 300]$

Dataset	Model	AUC	F1	Acc
DHA	Intrinsic	0.692	0.322	0.788
	NbrOnly-10	0.694	0.244	0.813
	Random-10	0.718	0.289	0.775
	Stacked-10	0.762	0.303	0.823
	NbrOnly-300	0.788	0.348	0.833
	Random-300	0.790	0.367	0.838
	Stacked-300	**0.800**	0.372	**0.844**
	AllNeighbors	0.793	**0.399**	**0.844**
SRMRB	Intrinsic	0.778	0.704	0.696
	NbrOnly-10	0.655	0.620	0.617
	Random-10	0.705	0.697	0.658
	Stacked-10	0.783	0.684	0.670
	NbrOnly-300	0.794	0.704	0.707
	Random-300	0.838	0.736	0.735
	Stacked-300	**0.840**	0.730	0.738
	AllNeighbors	0.828	**0.744**	**0.744**

4.3 Alternative Systems

Our proposed system described above, which we denote *Stacked-Ranked-k*, uses a similarity function to induce a subnetwork for the purpose of stacked inference. In order to test our hypotheses, we construct several alternative systems (see Table 3). In each case, we train the model on the same intrinsic and stacked features described above (as appropriate for that system). Where possible, we compare our results to several baselines from prior work [1, 9, 13, 14], as reported in a recent comparative study [12].

5. RESULTS

We discuss some differences in data imbalance between the two datasets and our choice of metrics, and our findings: using similar neighbors improve stacked inference, neighbors can provide good inference even without intrinsic features, and a stacked model outperforms existing classifiers.

Data Imbalance and Metrics. The results of our experiments are displayed in Table 4. Due to the unbalanced

[1] ciir.cs.umass.edu/downloads/

[2] https://dumps.wikimedia.org/

[3] https://github.com/dkpro/dkpro-jwpl

nature of the DHA dataset, neither F1 nor accuracy are representative metrics for classification. Thus, we focus most of our subsequent discussion on Area under ROC (AUC), a metric commonly used to evaluate unbalanced sets, as it is insensitive to dataset imbalance. We report F1 and accuracy results for comparison with prior work.

Similar Neighbors Improve Results, particularly for the first few neighbors. The predictive power of the stacked model grows with the number of neighbors. Results increase substantially within the first 25 neighbors, with diminishing returns afterwards. The Stacked classifier outperforms both the Intrinsic and Neighbor-only models, for both datasets and all metrics presented (see Table 4). For most values of k, our proposed system (which chooses neighbors according to a similarity metric), outperforms a random selection of the same number of neighbors, with the difference clearest when a small number of neighbors is used (figure omitted due to lack of space). As the number of neighbors increase and approach all neighbors of the page, the subnetwork approach converges to a "regular" stacked approach.

Neighbors Provide Quality Inference Without Intrinsic Features. As expected, each stacked model outperforms its equivalent Neighbors-only version, which ignores the intrinsic features of the page. Interestingly, in some cases the Neighbors-only model outperforms an intrinsic classifier (see Table 4), despite not receiving any features of the page itself. Further work is needed to examine this phenomenon.

Stacked Models Outperform Prior Work. There are some challenges in comparing our results to prior work on controversy detection in the SRMRB dataset, chief of which is that our results are reported on a more up-to-date Wikipedia dump (see [12] for comprehensive comparative analysis of controversy classification). Unfortunately, these results were reported only in terms of accuracy (percent correct) with no AUC or other metrics reported. With these constraints in mind, our result of 74.4% accuracy outperforms the Basic method (60%, [13]), the bipolarity method (56%, [1]), and the Mutual Reverts method (67%, [14]) - all results as reported in [12][4]. Our result of 74.4% is slightly lower than the Meta classifier [9] (75%)[5]. Notably, stacked models are ensemble methods and agnostic to the choice of intrinsic classifier for the problem, so any intrinsic classifier can be enhanced by applying our stacked classifier on top of it.

6. CONCLUSIONS AND FUTURE WORK

We present a novel stacked collective inference approach to detecting controversy in Wikipedia. By demonstrating that collective inference improves classification for this problem, we show that controversial articles exist in topical neighborhoods of controversy (i.e. exhibit homophily). Additionaly, we demonstrate that a subnetwork constructed based on similarity can yield better classification results than the default relationship in the dataset or randomly selected neighbors, particularly when a small subset of neighbors is used. This subnetwork approach can be generalized to other problem domains and is an effective way of incorporating similarity in collective and stacked inference. Depending on the degree of nodes and the tradeoffs between the computational cost of calculating pairwise similarity and those of running inference on all the neighbors, using similar neighbors may be preferable to all neighbors; we leave analysis of such tradeoffs to future work. The resulting stacked model improved over models using randomly selected neighbors, as well as over prior work. Future improvements in intrinsic classification of controversy can translate to additional improvements in the stacked model. Future work in collective classification could explore other similarity constructions. Automated detection of controversy holds promise for increased civic participation and a better informed public, by raising awareness and encouraging search users to consider alternative perspectives.

[4]We do not compare to the Editor Collaboration classifier [12], since it has intractable running time complexity (Sepehri Rad, personal communication) and cannot be reliably reproduced.

[5]Our Intrinsic classifier at 69.6% accuracy is the Meta classifier [9] without Talk Page features. While these features may be useful, Talk pages are infrequently used in non-English Wikipedias [14]. Using those features would likely improve the stacked model.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF grant number 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.

The authors thank Hosein Azarbonyad, Laura Dietz, Cibele Freire, Myungha Jang, Sandeep Kalra, Ariel Levavi, Raelen Recto, and Randy West for their comments on various drafts. Special thanks also go to Laura Dietz and Katerina Marazopoulou for fruitful discussions and to Hoda Sepehri Rad and Taha Yasseri for providing valuable resources.

7. REFERENCES

[1] U. Brandes, P. Kenis, J. Lerner, and D. van Raaij. Network analysis of collaboration structure in Wikipedia. *WWW*, 2009.

[2] M. Bröcheler, L. Mihalkova, and L. Getoor. Probabilistic Similarity Logic. *CoRR*, 2012.

[3] S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. 1998.

[4] S. Das, A. Lavoie, and M. Magdon-Ismail. Manipulation Among the Arbiters of Collective Intelligence. *CIKM*, 2013.

[5] S. Dori-Hacohen and J. Allan. Detecting controversy on the web. In *CIKM*, 2013.

[6] S. Dori-Hacohen, E. Yom-Tov, and J. Allan. Navigating Controversy as a Complex Search Task. In *Supporting Complex Search Tasks*, 2015.

[7] A. Fast and D. Jensen. Why stacked models perform effective collective classification. In *ICDM*, 2008.

[8] R. Jesus, M. Schwartz, and S. Lehmann. Bipartite networks of Wikipedia's articles and authors - a meso-level approach. *WikiSym*, 2009.

[9] A. Kittur, B. Suh, B. A. Pendleton, E. H. Chi, L. Angeles, and P. Alto. He Says, She Says: Conflict and Coordination in Wikipedia. In *CHI*, 2007.

[10] Z. Kou and W. W. Cohen. Stacked Graphical Models for Efficient Inference in Markov Random Fields. *SDM*, 2007.

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

[12] H. Sepehri Rad and D. Barbosa. Identifying controversial articles in Wikipedia. *WikiSym*, 2012.

[13] B.-q. Vuong, E.-p. Lim, A. Sun, M.-T. Le, H. W. Lauw, and K. Chang. On ranking controversies in Wikipedia: models and evaluation. In *WSDM*, 2008.

[14] T. Yasseri, R. Sumi, A. Rung, A. Kornai, and J. Kertész. Dynamics of conflicts in Wikipedia. *PloS one*, 2012.