Unsupervised Explainable Controversy Detection from Online News

Youngwoo Kim and James Allan

Center for Intelligent Information Retrieval
College of Information and Computer Sciences
University of Massachusetts Amherst
{youngwookim, allan}@cs.umass.edu

Abstract. Alerting users that a web page is controversial has been proposed as one method to support critical thinking about text and discourse. We propose an approach to discover controversial topics in a generic document using unsupervised training. Our approach comprises iterative training of a controversy classifier using a disagreement signal within comments and explaining the controversy of the document by generating a topic phrase describing it. Experiments show the effectiveness of our proposed training method using an EM algorithm. When controversial topic extraction is restricted to quality phrases and incorporates TextRank signals, it outperforms several baseline approaches.

Keywords: Controversy · Topic Extraction · Controversy Detection

1 Introduction

While search engines and social media are applauded for serving as effective information sources, they are also harshly criticized for delivering unverified and potentially harmful misinformation [14]. As an attempt to minimize such pitfalls, researchers have investigated controversy in the Web to predict misinformation and minimize the risk of it [18].

Much work on controversy has relied on certain signals available from the structure of the Web source, such as hashtags in social media which group information of similar content and thus contain inherent topic annotation [5, 11, 12, 19–21]. However, for a generic document without implicit or explicit topic annotations, the same approaches cannot be used. Even when it can be used, it is difficult for proposed systems to identify *what* is controversial [3, 7–9, 13].

Previous work on controversy detection on a generic document has two limitations. First, it relies on the topics that are labeled as controversial in Wikipedia [7, 8], or it relies on supervised human annotations [3], thus they may not be applicable to newly emerging topics. Second, it did not investigate how to provide an explanation for any controversy in the documents.

Our contributions are the following: First, we propose an unsupervised approach to build a controversy classifier using disagreement expressions. We aim

to detect topics that are expected to generate debates with numerous disagreements, which we view as the controversy in the news media. We show that a single feature of disagreement expression in the comment is enough to build an article-content-based controversy classifier without supervised training and we propose an EM algorithm to improve the training process.

Second, we propose a method to explain which topic is controversial in the document using the content-based controversy classifier. The controversy is explained by generating the phrases that describe the controversial topic in the document. We show that the quality of generated topic phrases can be improved by quality keyphrase constraints and a keyword extraction technique.

2 Unsupervised Controversy Classification

We target online news documents that contain both the article content and users' comments about the article. We thus define a "document" to be the pair of an article's content and its comment thread, though our goal will be to train a classifier that depends only on article content. To tackle this problem, we note that if a person were asked to decide what is controversial, one way might be to observe people's reactions to the article to get a sense which topics tend to generate more controversial debates. Following Beelen et al.[3], we use the presence of disagreement expressions to recognize controversy within comments. We use a text classification approach to find such expressions (section 2.2). However, because of likely errors in automatic detection we observed that disagreement expressions in a single document alone are insufficient to predict controversy. Thus, we decided to use disagreement in *comments* as a weak signal to train an article content classifier. We further improve this approach by re-training the comment text classifier using the article content classifier and iterating that process. This strategy is an instance of the EM algorithm [6, 10, 16].

2.1 EM algorithm for Controversy Classifier

We build two Language Model classifiers, where one is for an article content and the second is for comments (section 2.3).

Step 1 For a document x_i, the comment classifier f_c predicts whether the document is controversial, z_i, with

$$z_i = f_c(\theta_c^{(1)}, x_i) \tag{1}$$

where $\theta_c^{(1)}$ is the first set of parameters for the comment classifier. Based on the comment classifier's predictions, we assign a binary label z_i to every document in our corpus. Then the label z_i is used to get the article-content classifier's parameter set $\theta_a^{(1)}$.

$$\theta_a^{(1)} = \arg\max_{\theta} \sum_i z_i f_a(\theta, x_i) \tag{2}$$

Step 2 Again we predict each document's label using article content classifier f_a, and based on that label, get new parameters $\theta_c^{(2)}$.

$$z_i = f_a(\theta_a^1, x_i) \tag{3}$$

$$\theta_c^{(2)} = \arg\max_{\theta} \sum_i z_i f_c(\theta, x_i) \tag{4}$$

These two steps are iterated until convergence. For the controversy language model, equation 2 and equation 4 actually update $P(w|L_C)$ and $P(w|L_{NC})$.

2.2 Initial Signal

As an initial settings of z_i, a document is assigned a pseudo-label if it has more than certain number of disagreement in its comments. To estimate the number of disagreement expressions in the comments, we trained a Convolutional Neural Network based classifier using Authority and the Alignment in Wikipedia Discussions (AAWD) corpus [4, 15]. We take the first 100 comments as input and predict the number of disagreement in them. If the number of disagreement is larger than a threshold, the document is classified as controversial. We assumed the prior probability of a document having a 'controversy' label to be 0.5 and determined the threshold based on target corpus – i.e., such that half of the documents are (pseudo) controversial.

2.3 Language Model

As our primary controversy classifier model, we used the Controversy Language Model [13] which predicts a controversy ($D_C = 1$) by comparing whether the document is more likely to appear in a controversial document collection (L_C) or in a non-controversial document collection (L_{NC}).

$$\log P(D_C = 1) = \sum_{w \in D} \log P(w|L_C) - \log P(w|L_{NC}) \tag{5}$$

$P(w|L_C)$ and $P(w|L_{NC})$ are the probability of a word w in the collection of controversial documents and non-controversial documents. A document is classified controversial if $\log P(D_C) > T$ where T is set by a training corpus.

We also considered neural classifiers such as Convolutional Neural Network based text classifiers, but they turned out to be too unstable to be trained using the EM algorithm and also never outperformed the Language Model classifier.

3 Controversy Detection Explaination

When the trained classifier detects a controversial document, users will also expect an explanation of which topic in the document is actually controversial. We

choose to generate topic phrase which can predict that explanation. We do this by analyzing each document token's contribution to the classification decision and generating a topic phrase based on the contribution information. First, we describe restrict candidate phrases to those meeting standard of reasonableness, so that the user can clearly understand what the output phrase implies. Then, we explain how the contribution to the classification is evaluated and how it is transformed to score each candidate phrase.

3.1 Quality phrase as a candidate topic

Candidate topic phrases are restricted to be quality phrases that can be extracted from the target document. An n-gram is considered a quality phrase if (1) its document frequency exceeds a minimum, (2) it does not begin or end with stopword and (3) for the i$^{\text{th}}$ word w_i in the phrase, $P(w_i|w_{1:i-1}) > \lambda \cdot P(w_i)$. We used minimum document frequency=4, λ=10, and phrase length $n \leq 3$.

3.2 Candidates scoring

Candidate phrases are scored based on the degree to which phrase can explain the classifier's decision. The phrase with highest score is presented as the final output. Each token in the document is assigned a **contribution score** which represents how much it contributes to the classifier decision. For the Controversy Language Model, the contribution of each word is given as $R_w = \log P(w|L_C) - \log P(w|L_{NC})$. For neural classifiers, input contribution can be evaluated using contribution analysis techniques [1, 2]. A phrase's contribution score is sum of its terms' contributions, while each term's contribution is summed over all occurrences of the term in the document.

We added keyword scoring technique TextRank [17] as a **contribution independent score** for topic phrases. While the contribution to the classification decision is the most important factor in ranking explanations, we want the selected explanation to be representative and summarize other factors as well. TextRank score of the phrase is multiplied to the contribution score to achieve a final score of the candidate phrase.

4 Experiments

We present two experiments. The first experiment demonstrates the trained classifier's ability to correctly identify controversial documents. The second experiment evaluates how well the topic phrases generated from the classifier match human generated phrases. A qualitative analysis shows characteristics of topic phrases extracted from real world data.

For training, we collected unlabeled news documents from the Guardian, a British daily newspaper. We crawled the articles written in 2016 along with the related comments, which resulted in 66,763 news articles and 7,803,440 comments. We refer this collection as *Guardian16*.

Table 1. Controversy Classification Accuracy
Differences between upper three methods are statistically significant under $p < 0.05$

Method	**Accuracy**
Weak Signal	0.541
LM - Single Iteration	0.704
LM - EM	**0.746**
LM - Supervised	0.749
Human Annotator	0.744

4.1 Evaluation for classification

The model is trained using the *Guardian16* corpus. Part of the articles were labeled using Amazon Mechanical Turk. 6 annotators were asked if each document is about controversial topic or not. Documents with more than 3 'controversial' annotations were assigned final controversial label. which resulted in 281 controversial and 439 non-controversial documents.

Table 1 shows the controversy classification accuracy of various methods. The 'Weak Signal' classifier is based on the number of disagreement in the comments, which was our initial label. 'LM - Single Iteration' is the controversy language model trained by 'Weak signal' without additional iteration. 'LM - EM' is our proposed method. 'LM - Supervised' is the Language Model trained in a supervised setting in which 2/3 of the 720 documents were taken as training data and remaining were regarded as test data. Three splits were made and results were averaged to get the final accuracy. 'Human Annotator' is the hypothetical classification accuracy in which one of annotator's prediction is compared against the others. Note that all 'LM' methods classify based on the article contents alone (i.e., no comment).

4.2 Evaluation for explanation

Here, we evaluate our method's ability to predict the topic of the controversy. We collected 124 articles from `iSideWith.com`, which has a manually curated collection of the articles about a number of politically controversial topics. Those articles are from 13 controversial topics. Note that all of these documents are implicitly labeled as controversial. These documents are annotated with human generated topic phrases, which we adopt as ground truth for explanation and compared with output phrases from our methods.

Table 2 shows the evaluation for system generated topic phrases. As we have only one "gold" phrase, evaluation is using MRR, the reciprocal of the answer phrase's rank. Explanations generated with the conditions outperform the baseline group, which select phrases from any N-Gram ($N \leq 3$)

4.3 Quantitative analysis

For qualitative analysis we analyzed which topics are most controversial in the collection, which we generated by accumulating individual document's controver-

Table 2. Explanation Performance
Superscripts indicate the specified method is superior over numbered method ($p < 0.05$). Statistical significance was only measured between methods in the same group or the same model.

Restriction	#	Method	P@1	MRR
N-Gram	1	LM-EM	0.10	0.19
Quality Phrases	2	Random	0.06	0.12
	3	First N Phrase	0.15^{2}	0.21^{2}
	4	LM-EM	$0.26^{1,2,3}$	$0.38^{1,2,3}$
+ TextRank	5	TextRank only	0.24	0.36
	6	LM-EM	0.33	0.41

Table 3. Top controversial topics in the collection.

Rank	TextRank Only	**Proposed**
1	Trump	Trump
2	women	EU
3	EU	government
4	people	labour
5	police	tax
6	min	Clinton
7	mental health	party
8	children	prime minister
9	labour	climate change
10	tax	Corbyn

sial topic scores. For each document, the candidate topic phrases are assigned scores as explained in section 3. Then the phrase scores from each document are summed to get final controversial topics at the collection level. We used 'TextRank Only' as a baseline method. Table 3 shows the top topics extracted from *Guardian16*. TextRank captures less-controversial topics such as 'women', 'people' and 'children'. In contrast, the proposed method captures clearly controversial topics as top entries.

5 Conclusions

We introduced a classifier driven approach to detect controversial topics in the news media. We showed that the EM algorithm can improve the training process and the adding quality phrase information and keyword scoring helps to generate human friendly explanation from the classifier. As future work, we expect to extend disagreement signal driven controversy detection to generic web pages outside the news media, and to generate explanations in a detail-rich format.

Acknowledgment This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF grant #IIS-1813662. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor. We thank Kaspar Beelen for sharing the labeled data.

Bibliography

[1] Ancona, M., Ceolini, E., Oztireli, C., Gross, M.: Towards better understanding of gradient-based attribution methods for deep neural networks. In: 6th International Conference on Learning Representations (ICLR 2018) (2018)

[2] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one **10**(7), e0130140 (2015)

[3] Beelen, K., Kanoulas, E., van de Velde, B.: Detecting controversies in online news media. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1069–1072. ACM, New York, NY, USA. https://doi.org/10.1145/3077136.3080723

[4] Bender, E.M., Morgan, J.T., Oxley, M., Zachry, M., Hutchinson, B., Marin, A., Zhang, B., Ostendorf, M.: Annotating social acts: Authority claims and alignment moves in wikipedia talk pages. In: Proceedings of the Workshop on Languages in Social Media. pp. 48–57. Association for Computational Linguistics (2011)

[5] De Clercq, O., Hertling, S., Hoste, V., Ponzetto, S.P., Paulheim, H.: Identifying disputed topics in the news pp. 32–43 (2014)

[6] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society. Series B (methodological) pp. 1–38 (1977)

[7] Dori-Hacohen, S., Allan, J.: Detecting controversy on the web. In: Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. pp. 1845–1848. ACM (2013)

[8] Dori-Hacohen, S., Allan, J.: Automated controversy detection on the web. In: European Conference on Information Retrieval. pp. 423–434. Springer (2015). https://doi.org/10.1007/978-3-319-16354-3_46

[9] Dori-Hacohen, S., Jensen, D., Allan, J.: Controversy detection in wikipedia using collective classification. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 797–800. SIGIR '16, ACM, New York, NY, USA (2016). https://doi.org/10.1145/2911451.2914745

[10] Fessler, J.A., Hero, A.O.: Space-alternating generalized expectation-maximization algorithm. IEEE Transactions on Signal Processing **42**(10), 2664–2677 (1994). https://doi.org/10.1109/78.324732

[11] Garimella, K., Morales, G.D.F., Gionis, A., Mathioudakis, M.: Quantifying controversy on social media. Trans. Soc. Comput. **1**(1), 3:1–3:27 (Jan 2018). https://doi.org/10.1145/3140565

[12] Jang, M., Allan, J.: Explaining controversy on social media via stance summarization. In: The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1221–1224. ACM, New York, NY, USA. https://doi.org/10.1145/3209978.3210143

[13] Jang, M., Foley, J., Dori-Hacohen, S., Allan, J.: Probabilistic approaches to controversy detection. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. pp. 2069–2072. CIKM '16, ACM, New York, NY, USA (2016). https://doi.org/10.1145/2983323.2983911

[14] Kata, A.: A postmodern pandora's box: anti-vaccination misinformation on the internet. Vaccine **28**(7), 1709–1716 (2010). https://doi.org/10.1016/j.vaccine.2009.12.022

[15] Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)

[16] Mann, G.S., McCallum, A.: Generalized expectation criteria for semi-supervised learning with weakly labeled data. Journal of machine learning research **11**(Feb), 955–984 (2010)

[17] Mihalcea, R., Tarau, P.: Textrank: Bringing order into texts. In: Proceedings of the 2004 conference on empirical methods in natural language processing. pp. 404–411. Association for Computational Linguistics, Barcelona, Spain (July 2004)

[18] Qazvinian, V., Rosengren, E., Radev, D.R., Mei, Q.: Rumor has it: Identifying misinformation in microblogs. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1589–1599. Association for Computational Linguistics (2011)

[19] Yamamoto, Y.: Disputed sentence suggestion towards credibility-oriented web search. In: Proceedings of the 14th Asia-Pacific International Conference on Web Technologies and Applications. pp. 34–45. APWeb'12, Springer-Verlag, Berlin, Heidelberg (2012). https://doi.org/10.1007/978-3-642-29253-8_4

[20] Yasseri, T., Sumi, R., Rung, A., Kornai, A., Kertész, J.: Dynamics of conflicts in wikipedia. PloS one **7**(6), e38869 (2012)

[21] Zielinski, K., Nielek, R., Wierzbicki, A., Jatowt, A.: Computing controversy: Formal model and algorithms for detecting controversy on wikipedia and in search queries. Information Processing & Management **54**(1), 14–36 (2018). https://doi.org/10.1016/j.ipm.2017.08.005