Detecting Controversies in Online News Media

Kaspar Beelen
Informatics Institute
University of Amsterdam
k.beelen@uva.nl

Evangelos Kanoulas
Informatics Institute
University of Amsterdam
e.kanoulas@uva.nl

Bob van de Velde
Informatics Institute
University of Amsterdam
r.n.vandevelde@uva.nl

ABSTRACT

This paper sets out to detect controversial news reports using online discussions as a source of information. We define controversy as a public discussion that divides society and demonstrate that a content and stylometric analysis of these debates yields useful signals for extracting disputed news items. Moreover, we argue that a debate-based approach could produce more generic models, since the discussion architectures we exploit to measure controversy occur on many different platforms.

KEYWORDS

Controversy Detection, Media Analysis, Behavioral Analysis

ACM Reference format:
Kaspar Beelen, Evangelos Kanoulas, and Bob van de Velde. 2017. Detecting Controversies in Online News Media. In *Proceedings of SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan,* 4 pages.
DOI: http://dx.doi.org/10.1145/3077136.3080723

1 INTRODUCTION

With the advent of Web 2.0, the online world has become an intrinsic part of the public sphere. Growing interactivity and connectivity transformed the Web into a digital forum where discussions develop and societal disagreements arise. Controversies that divide public opinion are increasingly fought in the digital realm and with digital means.

Recognizing controversy is difficult, for algorithms as well as humans. In this paper, we develop a "hybrid" approach, combining insights from both social and computer science: first we determine key concepts coined by social scientists, and subsequently translate these to a generic but nonetheless predictive model of controversy. Instead of relying on platform-specific content or features, we argue that a discussion-based approach could yield a more widely applicable model for monitoring online disputes.

2 RELATED WORK

2.1 Controversy in the Social Sciences

Controversies exist as a type of **public debate**: they touch on issues that divide large segments of society [2, 9]. They emerge through the interaction between core-campaigners and broader sections of the public (termed occasional campaigners and sympathizers) [9]. Because they bear upon deeper rooted ideological divisions or opposing value systems, controversies tend to be unsolvable and **persist over time**. The increasing delineation of opposing views results in an ever widening disagreement or **polarization** [15]. Given that disputes flow from the participants' beliefs and values, the exchange of opinions is not limited to "facts", but invites strong **emotions** as well [8]. More linguistically inspired scholars such as Clarke [2] emphasized the **indexical** function of the term, pointing out how producers of discourse construct controversy by strategically naming and classifying events. Moreover, social psychologists [1], have argued that mental states of interlocutors are reflected in their linguistic style, implying that discussions on controversial topics may exhibit divergent stylometric patterns (i.e. a distinct **debating style**).

2.2 Controversy in Computer Science

Computer scientists, through coincidence or serendipity, concentrated on similar aspects when modeling and detecting online controversies. Debate structure, for example, plays a prominent role in Garimella et al. [5] who elicit public disagreement through "conversation graphs", a network constructed from tweets on a hashtag. Emotions have proven a powerful indicator as well. Popescu and Pennacchiotti [11] studied how controversial events develop on Twitter. They captured the level of polarization by computing how mixed the audience's response was in terms of sentiment. Perceiving controversies as primarily indexical, other studies relied on "Controversy Lexicons" to interrogate their data. Mejova et al. [10] analyze news reports using a crowd-sourced lexicon containing frequent content words for which participants were asked whether they signaled controversy or not. Also, Jang et al. [7] assessed the power of lexicon-derived features, building on the work of Cramer [3]. Roitman et al. [13] apply a manually crafted lexicon to retrieve controversial claims. Besides these feature types, Wikipedia counts as a crucial instrument for controversy detection. Previous research has leveraged the metadata associated with Wikipedia pages—the length of the discussion page, the presence of edits and reverts—to model dispute. Focusing on "editorial wars" Yasseri et al. [17], revealed the "dynamics of conflict" that lay behind the encyclopedia. Also in Dori-Hacohen and Allan [4] Wikipedia was adopted as a yardstick of controversy. Using a nearest neighbor approach they mapped Web pages to their closest Wikipedia articles–assuming that a site is controversial if the Wikipedia neighbors are. Our approach emphasizes the style *and* content of online conversations; it provides a generic method for detecting controversy not just based on *what* users discuss, but also *how* they perform the debate. Similar to Siersdorfer et al. [14] we apply controversy detection to news content. But whereas Siersdorfer et al. [14] focused on detecting controversial comments based on textual features (or polarizing

SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan

978-1-4503-5022-8/17/08…$15.00
DOI: http://dx.doi.org/10.1145/3077136.3080723

content based on rate divergence, i.e. the extent to which items receive likes and dislikes at similar rates), we attempt to classify articles by gauging a broader set of features.

3 DETECTING CONTROVERSY

3.1 Research Questions

The aim of this paper is to propose a debate-based method for detecting disputed content. To demonstrate our method, we look at comment threads associated with news articles—but the model applies to other contexts and platforms, as long as the discussion can be transformed to a post-reply tree. Our emphasis on online debates as a source of information, was driven by the scarcity of generic and adaptive approaches. The state-of-the-art relies heavily on platform-specific content (e.g. Wikipedia articles) or features (e.g. retweets on Twitter)—while discursive exchanges between users, the focus of this study, are found everywhere online. The paragraphs below demonstrate how monitoring discussions helps detecting controversial content by answering the following research questions: **RQ1** How to detect controversial newspaper articles based on their surrounding discussions? Which features prove most informative? **RQ2** How does this approach compare to other relevant baselines? Can different models be combined to improve accuracy?

3.2 Data Selection and Annotation

The data was sourced from the *theguardian.com*, the online version of the British broadsheet. According to the National Readership Survey, *guardian.com* ranks third in terms of popularity in the UK, just after the online editions of the Daily Mail and the Daily Mirror.[1] As the website content is freely accessible, the Guardian attracts a wide and ideologically diverse readership. This is reflected in the variety of opinions articulated by readers in the articles' comments section, which makes this platform an ideal location for monitoring disputed news. Using the Guardian API, we scraped all articles and their associated comments, published between September and November 2017, and selected a sample of 900 for manual annotation. We organized a crowd-sourcing task in which participants were asked to rate each article as either clearly non-controversial, possibly non-controversial, possibly controversial and clearly controversial. The labels were converted to an integer scale, from 1 (clearly non-controversial) to 4 (clearly controversial). For each article we obtained three annotations. Those with an average higher than 2.5 were categorized as "controversial". Using this cut-off, the annotated corpus split almost evenly into controversial and non-controversial articles.

3.3 Feature Space

The features extracted from the comments draw on different sources of information. **Linguistic Aspects**: These features capture linguistic variation between debates by counting the Part-of-Speech tags. **Structural Features**: Features that measure the formal aspects of the debate, such as the number of comments, the speed at which they were posted, and the percentage of replies. **Lexicons**: Instead of creating a hand-crafted or crowd-sourced lexicon, we chose to automatically generate an "agreement" and "disagreement" word list [12]. Starting with a list of manually selected seed-words that unambiguously mark agreement or its antonym [2] we extracted related words from embeddings trained on the Google News Corpus (referred to as V below). For each word w_i in V we computed a Lexicon score: $l_i = \sum_{j=1}^{k} cos(\boldsymbol{v}_i, \boldsymbol{v}_j)$, with v_i and v_j being the vector representation of the word w_i and the seed word w_j respectively. Consequently, we ranked all words by their l_i scores from high to low and selected the first 1000. [3] **Emotion**: Sentiment detection was performed with SentiStrength, a tool which has proven to obtain human-level accuracy on short texts. For each comment it produces a score between -4 (very negative) or +4 (very positive). Texts with a sentiment score between -1 and +1 were classified as neutral. To estimate how mixed the response was to an article, we followed the formula proposed by Popescu and Pennacchiotti [11], with $\#C_{emo}$ referring to the number of comments with sentiment orientation *emo*:

$$\frac{Min(\#C_{pos}, \#C_{neg})}{Max(\#C_{pos}, \#C_{neg})} \cdot \frac{\#C_{pos} + \#C_{neg}}{\#C_{pos} + \#C_{neg} + \#C_{neut}}$$

Other Features The WikiPedia Score WS_j of a given article a_j—or concatenation of the comments appended to that article—is defined as the sum of the cosine similarities of the Tf-Idf representation of the article (or comments) and the Tf-Idf vector of the pages listed as controversial on Wikipedia.

4 RESULTS

4.1 Comment-based Models

To answer **RQ1** we assess how accurately subsets of the comment-based feature space (see Table 1) predict the controversiality of a news report. We tested different models but opted for Random Forests (RF) and Support Vector Machines (SVM). The tables below show scores produced by Random Forests—which scored slightly better—with the exception of Table 3 which reports the weights of a SVM with linear kernel after training. Table 2 shows accuracy, f1 and precision scores obtained after 5-fold cross-validation by feature group: linguistic, structural, emotional, controversy and combined. The linguistic characteristics don't perform well, with only the structural features faring worse: opposed to our initial expectations, the size of the debate—expressed by the number of comments, or the rate at which they were posted—serves as a weak predictor. Emotion slightly outclasses the linguistic and structural subsets, with an accuracy of 70 per cent. The features we explicitly designed to capture the "controversial" aspects of a discussion work truly better, obtaining an accuracy of 75 per cent and a precision outclassing all previous models. Combining the feature sets improves the performance, irrespective of the chosen metric.

[1] See: http://www.nrs.co.uk/latest-results/nrs-padd-results/mobile/

[2] The seeds word list comprises words which are related to "disagreement" or "agreement" according to http://www.thesaurus.com/browse/agreement.

[3] Because antonyms are often closely located to each other in the vector space—the vector representation of "good" lays near to "bad"—"disagreement" words sometimes rank quite highly in the Agreement Lexicon. To filter out this noise, we discarded word w_i from lexicon L_{agr} if it happened to have a higher rank in lexicon L_{disa} (and vice versa).

[4] To measure offensive language we used the lexicon provided by Luis von Ahn http://www.cs.cmu.edu/ biglou/resources/bad-words.txt

[5] https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues

Table 1: Feature Space of the Model

Feature Name	Social Metric	Description
LING-POS	Debate Style	Percentage of tokens that belong to the same Part-of-Speech category With P-o-S either NN (noun), PR (pronoun), RB (adverb), JJ (adjectif), MD (modal), UM (interjection)
LING-QU	Debate Style	Percentage of tokens that are quotation marks
LING-LENGTH	Debate Style	Average (or variance) number of tokens per comment
LING-OVERLAP	Debate Style	Average (or variance) number of overlapping tokens between a post and a reply
EMO-POS—NEG—NEUT	Emotion	Relative number of positive, negative or neutral comments
EMO-REP-NEG—POS	Emotion	Relative number of replies with negative or positive sentiment
EMO-REP-DIFF	Emotion	The mean of the differences between the sentiment score of a post and the sentiment score of a reply to that post.
STRUC-REP	Debate	Relative number of comments that are replies
STRUC-NUM	Debate	Absolute number of comments
STRUC-ONE	Debate	Absolute number of comments posted one hour after the article was published
STRUC-RATIO	Debate	Number of comments divided by the time (expressed in seconds) between the first and last comment
CONTRO-EMO-MIX	Polarization	Indicates how mixed the response is in terms of sentiment.
CONTRO-CONTRA	Polarization	Contradiction score C developed by Tsytsarau et al. [16].
CONTRO-LEX-DIS	Indexical	The probability that a word belongs to L_{dis}
CONTRO-LEX-AGR	Indexical	The probability that a word belongs to L_{agr}.
CONTRO-LEX-OFF	Emotion	The probability that a word is an "offensive" term [4]
CONTRO-ANTONYM	Polarization	Number of WordNet antonym pairs divided by the number of posts
CONTRO-CL	Polarization	The Silhouette score obtained after k-means clustering of comments by user (k=2).
CONTRO-WIKI-SCORE	Context/Time	Summed similarity of the newspaper article to the set of controversial Wikipedia articles [5].

Table 2: Accuracy for Comment-based Model

	ACC	F1	PREC
LING	0.69	0.71	0.65
STRUC	0.60	0.66	0.56
EMO	0.70	0.73	0.63
CONTRO	0.75	0.73	0.75
COMBINED	0.77	0.76	0.75

Table 3: Features Weights of SVM with Linear Kernel

Features Non-Contro	Weight	Features Contro	Weight
LING-PR	-0.57	CONTRO-LEX-DIS	0.52
LING-OVERL-MEAN	-0.31	CONTRO-WIKI-SC.	0.29
STRUC-REP	-0.24	LING-VB	0.26
EMO-POS	-0.21	CONTRO-LEX-OFF	0.22
LING-LENGTH-MEAN	-0.18	CONTRO-ANTON.	0.22
LING-JJ	-0.13	CONTRO-CONTRA	0.22
LING-NN	-0.04	EMO-REP-NEG	0.22
EMO-VAR	0.04	LING-UH	0.20
EMO-REP-DIFF	0.07	LING-MD	0.17
STRUC-NUM	0.07	EMO-NEG	0.15

Inspection of the feature weights shows that lexicon-based indicators (offensive words as well as those indexing disagreement) act as solid predictors of controversy (CONTRO-LEX-DIS, CONTRO-LEX-OFF). But besides strong language, discussants seem to convey negative emotions at higher rates (EMO-REP-NEG, EMO-NEG), and deploy a more adversarial vocabulary (CONTRO-ANTONYM).

Table 4: Accuracy for Content-based Model

Type	ACC	F1	PREC
COMBINED-COMMENTS	0.77	0.76	0.75
WIKI-ARTICLES	0.72	0.70	0.69
TFIDF-ARTICLES	0.75	0.73	0.73
TFIDF-COMMENTS	0.76	0.74	0.72

The presence of positive sentiments, on the other hand, pushes documents to the zero (non-controversial) class. These results confirm the findings of Mejova et al. [10], who reported a prevalence of negative framing in controversial newspaper articles. Their observation that disputed articles lack strongly emotional words, is only partially corroborated by Table 3, which shows that non-controversial issues are represented in a more positive tone. Exchanges about non-controversial articles tend to be longer and remain on topic—suggested by a higher ratio of overlapping tokens between comments and the replies they invite. Even though linguistic features fare poorly when taken in isolation, they do appear as crucial predictors: debates on non-controversial items exhibit a higher reliance on personal pronouns which suggests that participants give more attention to their "footing", i.e. the positioning of self and others as participants in a discursive event [6].

4.2 Content-based Models

To answer **RQ2** we compare the above method to content-based models. Table 4 compares the above results to other relevant baseline methods: a classifier trained on the Tf-Idf representation of the article content (or concatenated comments). WIKI-ARTICLES predicts controversiality based on the similarity of the article content

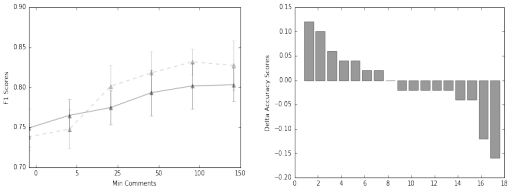

Figure 1: (Left) Comparison of the Comment-based (grey) and Tf-Idf model (black). The figure reports f1 scores (Y-axis) for different thresholds (N on the X-axis). (Right) Δ accuracy of Controversy vs Tf-Idf model. Each bar represents a batch of 50 documents, and results are sorted in descending order.

to each of Wikipedia's controversial pages. The content models were trained with a SVM.

Differences are small, and no clear candidate emerges on top of the others. However, it is remarkable that the simple debate-based model (remember it comprises only 29 features) generally performs better than all the others—including the Tf-Idf model which needs more than 15000 items to obtain similar accuracy.

Of course, our method depends—unfortunately but also self-evidently—on the presence of comments. If an article hardly generates any debate, the number of comments (the absence of user feedback) survives as the only predictor—one which was proven to contain a weak signal, see Table 3. However, if we exclude articles with less than N related posts, the performance increases, and clearly surpasses a content-based approach. The results are reported in Figure 1. Scores (mean and standard deviation) are obtained after training and testing on a randomized 66%-33% split, repeated fifty times. For sure, the gain in performance comes at a cost: the number of documents we can classify shrinks. But the predictions obtain higher precision, while simple content-based methods tend to remain stable. In short: for news items that spark a debate, looking at the discussion generally yields better results.

The fact that the models reported in Table 4 deliver similar results, does not imply they behave the same. Maybe these models capture different aspects and might complement each other? To assess if this is the case, we gauged how the classifiers performed on different subsets of the data. After cross-validation, we iterated step-wise (with the step size n set to 50) over the made predictions—an array with binary codes—and computed accuracy scores for all samples whose index fell within the range $\{n*(i-1)+1, n*i\}$ with i ranging from 1 to 18. For each batch of 50 documents we can thus compute the difference in accuracy (Δ) between two models. The sorted Δ scores reported in Figure 1 show that the difference between the comment and the content model (TFIDF-ARTICLES) is substantive: similarity in performance hides a difference in behavior.[6] The models fare better (or worse) on different parts of the corpus. To assess if the classifiers could complement each other, we created a stacked meta-learner, which builds a model on top of the predictions returned by the separate classifiers. However, the results obtained after 5-fold cross-validation show only a marginal improvement, as the meta-learner pushes the accuracy up to just 78 per cent.

[6] A comparison with WIKI-ARTICLES is not reproduced here, but the result was similar.

5 CONCLUSION AND FUTURE WORK

This paper outlined a novel strategy for detecting controversial news items. Defining controversy as a special type of debate, marked by polarizing dynamics and often charged with affect, we demonstrated that online discussions are an invaluable source of information: in most cases, a debate-based approach outperformed simple content baselines, and tended to fare consistently better when comments happen to be more abundantly available. Taken together, these observations suggest that analyzing online debates might serve a fruitful generic method for monitoring controversy. However, this short paper is just the thin edge of the wedge, a preliminary demonstration of a broader attempt to detect controversies on the Web. In future work, we aim to broaden and refine the notion of debate by including other Social Media platforms and distinguish between different types of participants who contribute to the controversy.

REFERENCES

[1] Cindy Chung and James W Pennebaker. 2007. The psychological functions of function words. *Social communication* (2007), 343–359.
[2] Adele E Clarke. 1990. Controversy and the development of reproductive sciences. *Social Problems* 37, 1 (1990), 18–37.
[3] Peter A Cramer. 2011. *Controversy as news discourse*. Vol. 19. Springer Science & Business Media.
[4] Shiri Dori-Hacohen and James Allan. 2013. Detecting controversy on the web. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, 1845–1848.
[5] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2015. Quantifying Controversy in Social Media. *CoRR* abs/1507.05224 (2015). http://arxiv.org/abs/1507.05224
[6] Erving Goffman. 1981. *Forms of talk*. University of Pennsylvania Press.
[7] Myungha Jang, John Foley, Shiri Dori-Hacohen, and James Allan. 2016. Probabilistic Approaches to Controversy Detection. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 2069–2072.
[8] Daniel J Levi and Elaine E Holder. 1988. Psychological factors in the nuclear power controversy. *Political psychology* (1988), 445–457.
[9] Brian Martin. 2014. *The Controversy Manual. A practical guide for understanding and participating in scientific and technological controversies*. Sparsns, Sweden, Reading, Massachusetts.
[10] Yelena Mejova, Amy X Zhang, Nicholas Diakopoulos, and Carlos Castillo. 2014. Controversy and sentiment in online news. *arXiv preprint arXiv:1409.8152* (2014).
[11] Ana-Maria Popescu and Marco Pennacchiotti. 2010. Detecting controversial events from twitter. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 1873–1876.
[12] Ludovic Rheault, Kaspar Beelen, Christopher Cochrane, and Graeme Hirst. 2016. Measuring Emotion in Parliamentary Debates with Automated Textual Analysis. *PloS one* 11, 12 (2016), e0168843.
[13] Haggai Roitman, Shay Hummel, Ella Rabinovich, Benjamin Sznajder, Noam Slonim, and Ehud Aharoni. 2016. On the retrieval of wikipedia articles containing claims on controversial topics. In *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee, 991–996.
[14] Stefan Siersdorfer, Sergiu Chelaru, Jose San Pedro, Ismail Sengor Altingovde, and Wolfgang Nejdl. 2014. Analyzing and mining comments and comment ratings on the social web. *ACM Transactions on the Web (TWEB)* 8, 3 (2014), 17.
[15] Sidney Tarrow. 2008. Polarization and convergence in academic controversies. *Theory and Society* 37, 6 (2008), 513–536.
[16] Mikalai Tsytsarau, Themis Palpanas, and Kerstin Denecke. 2010. Scalable discovery of contradictions on the web. In *Proceedings of the 19th international conference on World wide web*. ACM, 1195–1196.
[17] Taha Yasseri, Robert Sumi, András Rung, András Kornai, and János Kertész. 2012. Dynamics of conflicts in Wikipedia. *PloS one* 7, 6 (2012), e38869.