Detecting Controversies in Online News Media

Beelen, K.; Kanoulas, E.; van de Velde, R.

DOI
10.1145/3077136.3080723

Publication date
2017

Document Version
Final published version

Published in
SIGIR’17 : proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval

License
Article 25fa Dutch Copyright Act (https://www.openaccess.nl/en/in-the-netherlands/you-share-we-take-care)

Citation for published version (APA):
https://doi.org/10.1145/3077136.3080723

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 426, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)
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.vandevalde@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:
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 for 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 elicited 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
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. 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 wj in V we computed a Lexicon score: \( l_i = \sum_{j=1}^{k} \cos(v_i, v_j) \), 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 much the response was to an article, we followed the formula proposed by Popescu and Penachiotti [11], with \( C_{emo} \) referring to the number of comments with sentiment orientation emo:

\[
\text{Min}(C_{pos}, C_{neg}) = \frac{#C_{pos} + #C_{neg}}{#C_{pos} + #C_{neg} + #C_{neut}}
\]

Other Features: The Wikipedia Score \( W_S \) 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 controversy 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 The seeds word list comprises words which are related to “disagreement” or “agreement” according to http://www.thesaurus.com/browse/agreement.
2 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 \( w_j \) from lexicon \( L_{ag} \) if it happened to have a higher rank in lexicon \( L_{dis} \) (and vice versa).
3 To measure offensive language we used the lexicon provided by Luis von Ahn http://www.cs.cmu.edu/ biglou/resources/bad-words.txt
Table 1: Feature Space of the Model

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

Table 2: Accuracy for Comment-based Model

<table>
<thead>
<tr>
<th></th>
<th>ACC</th>
<th>F1</th>
<th>PREC</th>
</tr>
</thead>
<tbody>
<tr>
<td>LING</td>
<td>0.69</td>
<td>0.71</td>
<td>0.65</td>
</tr>
<tr>
<td>STRUC</td>
<td>0.60</td>
<td>0.66</td>
<td>0.56</td>
</tr>
<tr>
<td>EMO</td>
<td>0.70</td>
<td>0.73</td>
<td>0.63</td>
</tr>
<tr>
<td>CONTRO</td>
<td>0.75</td>
<td>0.73</td>
<td>0.75</td>
</tr>
<tr>
<td>COMBINED</td>
<td>0.77</td>
<td>0.76</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Table 3: Features Weights of SVM with Linear Kernel

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

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).

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.
with binary codes—and computed accuracy scores for all samples were trained with a SVM. (with the step size of 50 documents, and results are sorted in descending order. 6

a/f_ter 5-fold cross-validation show only a marginal improvement, as returned by the separate classi/f_iers. However, the results obtained stacked meta-learner, which builds a model on top of the predictions particles with less than

assess if the classi/f_iers could complement each other, we created a

similarity in performance hides a difference in behavior. [17] Taha Yasseri, Robert Sumi, Andr´as Rung, Andr´as Kornai, and J´anos Kert´esz. 2012. 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 \( \{ n \ast (i - 1) + 1, n \ast i \} \) with i ranging from 1 to 18. For each batch of 50 documents we can thus compute the difference in accuracy (\( \Delta \)) between two models. The sorted \( \Delta \) 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.

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 general 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


---

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