Technology Review: Controversy Detection

Author: Ryan March (ryanm14)

Detecting controversy in text is a valuable endeavor that provides many benefits. For example, detecting controversy on the web can help inform discourse on a large scale (such as nationwide or even worldwide). Identifying controversial topics has impact not just in political arenas – where spirited debates are crucial for evaluating appropriate legislation and policies – but also in a myriad of other venues such as entertainment and media. To demonstrate the utility of detecting controversy in subjects other than politics, imagine a production company that wishes to evaluate the polarity around a certain show to determine whether or not another season of it would be lucrative. Sometimes controversy could even be seen as desirable (if the adage "no publicity is bad publicity" is to be believed, at least in certain contexts)! However, controversy detection remains a difficult task for multiple reasons. First, society is dynamic, and the rubric for contentiousness is fluid and changing. Second, controversies are like ideologies in that they are latent instead of explicitly referenced. Finally, controversy spans a wide range of topics and is often accompanied by domain-specific vocabulary [1]. Since automatic mechanisms for detecting controversy prove to be both meaningful and challenging, researchers in the text mining space have developed multiple approaches to detect controversial topics. These approaches have had varying results, strengths, limitations, and applications.

There are three common approaches for a substantial portion of prior work: lexical approaches, explicit modeling, and matching models [1]. Lexical approaches leverage signal terms that indicate apparent controversy, usually via bag-of-word classifiers or lexicons/lexicon-based language models. Explicit modeling relies on platform-specific features, like a user-provided "controversial" tag on Wikipedia articles. Matching models combine these two approaches by attempting to identify lexical similarities between a reference dataset that maintains explicit features and an evaluation text [1]. Unfortunately, these three approaches tend to suffer from a lack of generalizability. Lexical methods will likely overfit the training set, memorizing the controversial topics instead of establishing a broader view of a "language of controversy" [1]. Meanwhile, matching models are inherently specific to the given platform they are trained on – for example, it is unlikely that a news article will have the same kind of "controversial" tag that a Wikipedia article might have [1].

The specific approach taken by Kim and Allan was able to achieve a more generalizable solution. The authors aimed to overcome two fundamental limitations. One such limitation is that many methods rely on supervised learning, requiring the dataset to either have built-in labels (i.e., topics labeled as controversial in Wikipedia) or to be enriched with human annotations. Another limitation is a lack of explainability in the results from controversy detection algorithms [2]. The authors used lexical methods to propose an unsupervised classifier based on disagreement expressions. By codifying these disagreement expressions, topics that generated debates leading to more disagreements could be identified. The authors achieved this with one feature around disagreement expression in the comments of news articles. Then, Expectation Maximization was leveraged to assist in training [2]. The explainability limitation was circumvented by generating phrases from a candidate topic phrase with a high contribution to the classification decision. These topic phrases were restricted to a certain quality so that a user could understand the output of the resulting explanation [2]. Kim and Allan cleverly combined several approaches to build a controversy classifier for news articles. The authors noticed that detecting controversy in just a single document was challenging. So they included user comments to identify disagreement as a weak signal to train their article content classifier [2]. To estimate the number of disagreement expressions in the comments, the authors trained a Convolutional Neural Network based on a known corpus and used that to evaluate the first 100 comments in a given news article. If the disagreement was above some threshold, an initial pseudo-label was assigned to the corresponding document [2].

In recent years, other authors have started focusing on social media as a hotbed for controversy. Twitter is a popular platform for this type of research because of its large user base, focus on reporting or reading about news and current events, and high concentration of debate-filled discussion [3]. From here, a natural extension to the prior research came from graphs. Garimella (et al.) focused on an approach to quantify controversy in any domain (not just specific domains that the authors had domain knowledge of) by using conversation graphs based on Twitter threads and discussions [4]. The authors evaluated various approaches that differed in their graph construction methodology and datasets. Overall, the research proposal employed a three-stage pipeline, comprised of graph building, graph partitioning, and controversy measure. The output of this pipeline was a quantity representing how controversial a topic is. Larger values of this quantity suggest higher degrees of controversy [4].

The goal of the graph construction phase of the pipeline was to build a conversation graph to represent activity about specific, individual topics. For this effort, topics were defined as related hashtags and social media activity (for example, posts) related to or matched that set of hashtags [4]. The authors innovated on the typical definition of "topic," which historically has specified that one hashtag defines one topic. Such a definition is often too limited, with the authors offering examples where this is too restrictive – for example, when two opposing viewpoints on a topic use different hashtags to represent their perspectives. Instead, the definition of a topic was given by a set of hashtags that co-occur with a specified seed hashtag, allowing the authors to define clusters [4]. Unfortunately, this also had drawbacks since some common hashtags co-occur with a large number of hashtags, even when the relationship between them is tenuous. To mitigate this, the authors tried to account for hashtag popularity by taking a document frequency of all hashtags (in a subset of posts) and normalizing them by an inverse document frequency. These steps allowed the authors to obtain the top-k most similar hashtags to the seed hashtag [4].

Once these hashtags were retrieved, the next step was creating the graph. The authors retrieved "items" (for example, posts) related to the particular hashtag. Each of these items also had a specific user associated with them. So, each user contributing to a particular topic was assigned to one vertex. Then, edges were constructed between the vertices to represent either endorsement, agreement, or a shared point of view between the involved users [4]. Different graph construction mechanisms were attempted for this phase, including a graph based on who users follow, what users retweeted, and what content was present in the posts [4].

The second phase of the pipeline was graph partitioning. This stage split the graph into exactly two partitions, with the authors noting that controversies involving more than two sides could be a subject for further study. Intuitively, by splitting the graph into two sections, the hope would be to extract two distinct "sides" or viewpoints of the related discussion. The authors describe this idea by suggesting that splitting users into two sides according to their point of view would allow them to determine what those two sides were [4] precisely. The final stage was quantifying the controversy given the constructed graph as input. For this, the authors again tried multiple approaches, such as random walks of the graph, betweenness centrality of the graph, and even low-dimensional embeddings, among others [4].

Subsequent work by Zarate (et al.) built upon this work to create a four-phase pipeline consisting of graph building, community identification, embedding, and controversy score computation [3]. Via this pipeline, the authors were able to extend previous approaches by including features extracted by techniques such as vocabulary analysis since different communities or topics may have various linguistic quirks associated with them [3]. Generally, this approach leveraged Garimella's work while extending it to improve accuracy and run-time efficiency. One significant improvement of Zarate's system is that it can extend beyond English, where the previous work utilized language-specific tools that do not work reliably for other languages [3].

Overall, controversy detection is a subject of active investigation, with new approaches developed to overcome limitations encountered by previous work. Some authors approach the problem with the purview of generality, hoping to achieve a classifier that works on standalone articles and web pages. Other authors focus on discussion-based forums like Twitter, recognizing that controversy may be easier to detect when multiple viewpoints are considered. Early results did seem successful in determining whether something was controversial but had limitations in explaining exactly which topics or sub-topics from which that controversy originated. Later, this limitation was addressed by using topic phrases to generate explanations of what exactly was considered controversial. In more recent years, with a focus on social media platforms, authors were able to leverage graph techniques to improve results and detect bifurcation in topics to quantify their disagreement or controversy. The field now has many possible extensions. There are possibilities to modify graph-building phases or similarity measures and opportunities to include various language-specific or language-agnostic features. Many researchers agree, however, that having automated detection of controversy without the need for domain-specific knowledge or supervised labels has the potential to impact the nature of discourse online and elsewhere significantly.

References

[1] Linmans, Jasper, Bob van de Velde, and Evangelos Kanoulas. "Improved and robust controversy detection in general web pages using semantic approaches under large scale conditions." *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 2018.
[2] Kim, Youngwoo, and James Allan. "Unsupervised explainable controversy detection from online news." *European Conference on Information Retrieval*. Springer, Cham, 2019.
[3] Zarate, Juan Manuel Ortiz de, et al. "Measuring controversy in social networks through nlp." *International Symposium on String Processing and Information Retrieval*. Springer, Cham, 2020.
[4] Garimella, Kiran, et al. "Quantifying controversy on social media." *ACM Transactions on Social Computing* 1.1 (2018): 1-27.