ControversialQA: Exploring Controversy in Question Answering

Zhen Wang[*1], **Peide Zhu**[*2], **Jie Yang**[†2]
[1] Tokyo Institute of Technology, Japan
[2]Delft University of Technology, Netherlands
wzh@lr.pi.titech.ac.jp, {p.zhu-1, j.yang-3}@tudelft.nl

Abstract

Controversy is widespread online. Previous studies mainly define controversy based on vague assumptions of its relation to sentiment such as hate speech and offensive words. This paper introduces the first question-answering dataset that defines content controversy by user perception, i.e., votes from plenty of users. It contains nearly 10K questions, and each question has a best answer and a most controversial answer. Experimental results reveal that controversy detection in question answering is essential and challenging, and there is no strong correlation between controversy and sentiment tasks. We also show that controversial answers and most acceptable answers cannot be distinguished by retrieval-based QA models, which may cause controversy issues. With these insights, we believe ControversialQA can inspire future research on controversy in QA systems. Code and dataset are available at https://github.com/zhenwangrs/CQA.

Keywords: Controversy, Question Answering, Sentiment Analysis

1. Introduction

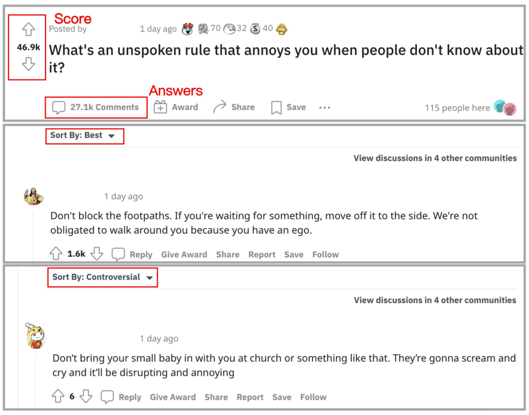
Large numbers of people are participating in online discussions on a daily basis, asking and answering questions, or expressing their opinions on certain topics, on various platforms such as Twitter, Facebook, and Reddit. It is common that the discussions may turn in opposite directions, and controversy may arise when people have conflicting opinions, especially on political, health or entertainment topics. With the development of social media, controversy now spreads over everywhere on the Web. Online platforms need to detect controversy in these discussions since the controversial texts may contain fake information, or some user groups may find them inappropriate, offensive, or unwanted.

Currently, there are only limited resources to investigate controversy on online discussion platforms; most of them are based on Twitter (Addawood and Bashir, 2016; Garimella et al., 2017a). Some researches regard debates over Twitter topics as the source of controversy (Addawood and Bashir, 2016; Garimella et al., 2017a). Garimella et al. (Garimella et al., 2017b) study the ebb and flow of controversial Twitter debates on certain topics. The research by Vilella et al. (Vilella et al., 2021) shows that disinformation is pervasively existing in debates. Lots of previous controversy-related researches assume controversy is strongly related to sentiment (Smith et al., 2013; Mejova et al., 2014). Early research on social media controversy focuses mainly on political activities, especially the president election, for the purpose of predicting the election result (Adamic and Glance, 2005; Conover et al., 2012). And the latter, Smith et al. (Smith et al., 2013) extend controversy research to a broader area, using a set of controversial topics on Twitter to investigate user behavior. They found that Twitter tends to relay one person's views to others who hold similar views. Mejove et al. (Mejova et al., 2014) find that there are more negative effects and biased language in the reader's discussions on controversial online news. While demonstrating the existence and the potential negative impact of controversy, these previous works define controversy using syntactic heuristics, i.e., characterizing controversy as a consequence of hate speech or offensive words. This is not necessarily true since the controversy is usually seen as a property of the semantics, e.g., the conflicting opinions among people, and sometimes can only be judged by user perception.

Since previous definitions are not exactly in line with the connotation of controversy, in this paper, we introduce ***ControversialQA***, the *first-of-its-kind* dataset collected focusing on controversial discussions, where controversy is defined by user perception (votes from plenty of users). We specifically focus on content in question answering (QA) since QA forums in Reddit tend to inspire discussions from people with various opinions; therefore, controversy results in both the question and the answer. Based on ControversialQA, we first propose a series of controversy detection tasks and show that it is a novel and challenging dataset. Second, in contrast to assumptions in previous work, we reveal that controversy detection is not firmly correlated to existing sentiment analysis or hate speech detection tasks. Furthermore, our experiments on ControversialQA prove that the current retrieval-based QA model may be opposed by controversy issues, that is, instead of providing the most accept-

[*] Equal Contribution. [†] Corresponding author.

LREC-COLING 2024, pages 3962–3966
20-25 May, 2024.



What's an unspoken rule that annoys you when people don't know about it?

Don't block the footpaths. If you're waiting for something, move off it to the side. We're not obligated to walk around you because you have an ego.

Don't bring your small baby in with you at church or something like that. They're gonna scream and cry and it'll be disrupting and annoying

Figure 1: An sample from r/AskReddit. The upmost two red boxes represent the question score and the current number of answers. The other two show a top ranked answer using *sort by best* and *sort by controversial*.

able answer, it may choose answers that can lead to opinion split.

2. Dataset

2.1. Data Collection

Since we focus on controversy detection in open online QA, we chose the *r/AskReddit* thread in Reddit [1] as the data source. *r/AskReddit* is a subreddit where users can ask questions or answer questions in other posts. With the help of API [2] provided by Reddit, we collect questions that have been posted to *r/AskReddit* over the past ten years. For each question, we collect the question itself and a most controversial answer; for comparison, we also collect a best answer which is supported by most of the readers. A sample can be found in Fig. 1. The first two red boxes represent the question score (the difference between upvotes and downvotes that users vote for this question) corresponding to the question quality and the current number of answers for the question.

The default sorting formula on Reddit is *sort by best*, where the larger the difference between the number of *upvotes* and *downvotes*, the higher the answer ranks. Another sorting formula is *sort by controversial*, where an answer with a higher controversial score will be listed ahead. The formula [3] Reddit uses to calculate the controversial score is shown in Equation 1.

Metric	Counts
#Total Instance	9,952
#Average tokens of question	15.58
#Average tokens of best answer	58.49
#Average tokens of controversial answer	52.99

Table 1: Statistics of ControversialQA.

$$\begin{aligned}\text{magnitude} &= ups + downs\\ \text{balance} &= \begin{cases}\frac{downs}{ups}, & ups > downs;\\ \frac{ups}{downs}, & ups \leq downs.\end{cases}\\ \text{controversial\_score} &= magnitude^{balance}\end{aligned} \tag{1}$$

where ups is the number of "upvote" and $downs$ is the number of "downvote". This algorithm ranks an answer to be more controversial when it has a large number of upvotes and downvotes.

2.2. Quality Control

To ensure the quality of the dataset, that is, to ensure that a controversial answer indeed receives upvotes and downvotes by a large number of users, we only retain questions with more than 100 answers and more than 100 question scores. Since the answers on Reddit sometimes come up with meaningless words or phrases such as "no" and "knock knock", only samples whose answers contain more than 20 words are retained. To make the distribution of the answer length more balanced, we also delete samples with too long answers if they contain more than 150 words. The statistics of ControversialQA are shown in Tab. 1.

3. Experiment and Analysis

3.1. Experimental Settings

We randomly split the dataset into training/validation/testing sets in the ratio of 8:1:1 for all the tasks. We conduct experiments on different tasks using BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) models provided by Hugginface, since these pretrained language models have achieved state-of-the-art results on various text classification tasks. To make the QA pairs conform to the input format of BERT, we use `[SEP]` to concatenate the answer and question to " `[CLS]` Answer `[SEP]` Question". The output of the `[CLS]` token is used for linear classification tasks. We use the F1 score as the evaluation metric. For each training process, we train the model with the training set, retain the model that performs best in the validation set, and apply it in the testing set. Adam (Kingma and Ba, 2014) is

[1] https://www.reddit.com/r/AskReddit/

[2] https://www.reddit.com/prefs/apps

[3] https://github.com/reddit-archive/reddit/blob/master/r2/r2/lib/db/_sorts.pyx

used as the optimizer, and the learning rate is 1e-5. Cross entropy loss is used as the loss function.

3.2. Controversy Detection on ControversialQA

We conduct two tasks on ControversialQA: ***(I) Controversy Classification Task:*** a QA pair is used as the input, and the output is a bool value representing whether this answer will cause controversy for the current question. ***(II) Controversy Selection Task:*** given two candidates, a model should correctly select which answer will cause controversy. The training process is the same as that of controversy classification. During the test, for each sample, we use softmax to calculate the score of the two answers, respectively, and take the one with the higher score as the controversial answer.

The experimental results are shown in the Tab. 2. Combining the two experiments, we get three major findings. (1) RoBERTa-large achieves the best performance in both tasks, which achieves the highest at 74.83 and 86.84 in F1, respectively, showing that current methods can achieve relatively good results, yet there is still a lot of room for improvement. (2) Directly detecting controversy of an answer is more difficult than comparing the controversy of two answers (**74.83** vs. **86.84**); (3) For both tasks, the large version of the model performs better than the base version, and the better pretrained model performs better (RoBERTa better than BERT), proving that parameters and pretraining are helpful to improve the accuracy in detecting controversy.

3.2.1. Case Study

In Tab. 3, we present three samples that are incorrectly classified in the controversy selection task. In the first question, the description of gift cards leads to controversy. Some people think gift cards are helpful because the cards allow more freedom to buy gifts, while some don't like gift cards because the cards force them to buy things in a store they don't like. In the second question, the judgment to *Breaking Bad Season 4* causes controversy. Some people think the season is interesting, while some others hold the opposite attitude as the one providing the given answer. As for the last question, many of the 9/11 conspiracy theories have already been disproved, but some people still believe in them.

Those samples reveal that controversy in the real world is caused by various factors such as personal experience, political opinions, etc., which can often go beyond the text itself. Therefore, to achieve better controversy detection, the model should consider not only the answer text, but also contextual, commonsense, or external knowledge.

Model	**AR**	**AM**	**AW**
Controversy Classification			
BERT-base	68.65	65.29	62.88
BERT-large	70.01	60.99	63.24
RoBERTa-base	71.91	67.21	65.17
+*sentiment*	68.51	66.20	62.44
+*offensive*	69.83	66.76	64.26
+*irony*	70.18	65.75	63.05
+*hate*	70.17	65.32	65.06
+*emotion*	70.51	66.50	62.78
RoBERTa-large	**74.83**	**69.96**	**66.50**
Controversy Selection			
BERT-base	78.51	72.12	70.19
BERT-large	80.12	74.01	72.02
RoBERTa-base	83.53	76.15	73.80
+*sentiment*	82.12	74.65	72.38
+*offensive*	80.92	74.26	71.69
+*irony*	80.82	73.94	71.15
+*hate*	81.32	74.57	73.14
+*emotion*	81.93	75.36	72.76
RoBERTa-large	**86.84**	**77.80**	**75.99**

Table 2: F1 score of controversy classification and choice. **AR** means *r/AskReddit*, **AM** means *r/AskMen*, **AW** mean *r/AskWomen* (Section 3.2.2).

3.2.2. Out of Domain Test

Since *r/AskReddit* is an open domain topic, we want to determine whether the model trained on it can be migrated to other specific domain topics. Therefore, we collect two auxiliary datasets with a similar size to the ControversialQA testing set from *r/AskMen* and *r/AskWomen* using the same collecting procedure and then apply the trained models to them. *r/AskMen* primarily aims at male Reddit users, asking questions about men and expecting men to answer them, while *r/AskWomen* is just the opposite. Experimental results are shown in the **AM** and **AW** columns of the Tab. 2.

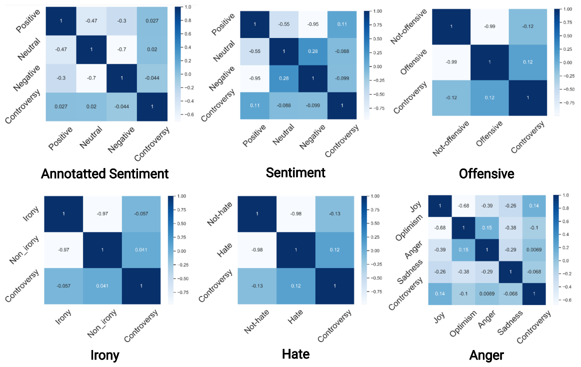
A huge gap between in- and out-of-domain performance can be observed. Notably, gender has a significant impact on discussion subjects and opinions. This shows that although *r/AskReddit* is an open-domain topic that contains a variety of questions involving all genders, applying the model directly to a specific domain does not yield the same outcome. Furthermore, this implies that controversy detection is closely related to user demographics.

3.3. Can Sentiment Help Detect Controversy?

Sentiment analysis is usually involved in previous controversy studies. To investigate whether the controversy detection task overlaps with sentiment analysis, first, instead of original RoBERTa, we perform controversy detection by fine-tuning TweetE-

Question	Choice A	Choice B
What is one thing people do NOT want for christmas?	Gift cards. Don't force me to spend virtual money at a store I don't shop at, only to have to fork over a few bucks at the register to keep there from being a few bucks left on the card. We prefer cash...	When we were younger my brother and I thought it would be funny as hell to give our little sister coal for Christmas ... of course she cried her eyes out so I'm gonna go with coal if you're a kid.
What TV series isn't worth finishing?	That 70s show. Generally people say that Randy was the reason for the shitty ending, but IMO the show shouldn't have tried to ...	I could never get through Breaking Bad. ive tried a few times and just cant. I cant get last season 4 it just gets so slow and boring
What is one conspiracy that you firmly believe in? and why?	9/11 was an inside job. If you examine the evidence for an inside job theory vs. the evidence available for the Bush administration's theory it is overwhelmingly in the favor of an inside job...	Cinnabon vents their oven exhaust directly into the food court to increase sales. Can you smell the food from Panda Express or McDonalds in the food court? Of course not...

Table 3: Three error predicted samples by fine-tuned RoBERTa-large. The choice with green color is the ground-truth answer that will cause controversy.



Annotatted Sentiment

Sentiment

Offensive

Irony

Hate

Anger

Figure 2: Correlation between **controversy** and different types of **sentiments**.

val (Barbieri et al., 2020), a collection of RoBERTa-base models pretrained on several different kinds of sentiment dataset, including *sentiment*, *offensive*, *irony*, *hate*, *emotion*, on ControversialQA. The results are shown in Tab. 2. Experimental results show that pretraining on sentiment datasets cannot help to improve the performance on controversy tasks, and the model with sentiment bias is instead even harmful in detecting the controversy.

To better reveal the relationship between controversy and sentiment, we then label QA pairs by both those sentiment models and humans. Human sentiment annotation is conducted with three experts and harvests 500 labeled pairs by simple major voting. Finally, we calculate their correlation with ground-truth controversy labels. The results are shown in Fig. 2. The first "Annotated Sentiment" map is calculated by the correlation between human annotated sentiment labels and ground-truth controversy labels. And the other five maps are calculated by the correlation between the labels generated by pretrained TweetEval models and ground-truth controversy labels. It can be observed that there is no strong correlation between those sentiments and controversy. Neither automatic nor human sentiment labeling shows a strong correlation between previous sentiment tasks and ControversialQA, which proves compared to previous sentiment-related tasks, our ControversialQA is a brand new task that worthy studied.

3.4. How Does Controversy Influence Retrieval-based QA Models?

Compared to controversial answers, which may contain fake information or views opposed by some user groups, the best answer should always be the first choice of IR-based QA model. To figure out whether current QA models can evaluate the quality of the answers from the controversial perspective, we use a DistilBert (Hofstätter et al., 2021) retriever pretrained on MS MARCO (Nguyen et al., 2016) to calculate the score between the question and two answers separately, where one is the best answer which is supported by the majority of users and the other is the controversial answer. We perform statistics across the entire dataset, and the result shows that there is a **52.43%** (ideal is 0%) chance that the model regards the controversial answer as more relevant (better) than the best one. This proves that current QA models that rank answers merely rely on the relevance between question and answers may raise controversy issues since they are not able to distinguish answer controversy. This can cause an answer that appears to fit the question, but actually contains some controversial information, such as fake news or gender discrimination, ranking at the top, which is unacceptable for real-world scenarios. Especially with the development of large language models (LLM), people are increasingly relying on content generated by LLM. There is now some researches focus on detecting the authenticity of answers generated by large models (Lai et al., 2024). We believe that our dataset can help facilitate research in this area and to improve the appropriateness and credibility of answers.

4. Conclusion

In this paper, we present ControversialQA, the first dataset used for controversy detection in QA. We then conducted comprehensive experiments on this dataset, which show that controversy detection is a challenging task. With a set of experiments, we first reveal that controversy has no strong correlation with sentiment tasks. We further prove that retrieval-based QA models cannot distinguish between the most acceptable answer and controversial answer semantically and may be opposed by controversy issues in real-world scenario. We believe these insights can inspire future research in several fields.

5. Bibliographical References

Lada A Adamic and Natalie Glance. 2005. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43.

Aseel Addawood and Masooda Bashir. 2016. “what is your evidence?” a study of controversial topics on social media. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 1–11.

Michael D Conover, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. 2012. Partisan asymmetries in online political activity. *EPJ Data science*, 1(1):1–19.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2017a. The effect of collective attention on controversial debates on social media.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2017b. The ebb and flow of controversial debates on social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. *arXiv preprint arXiv:2104.06967*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Zhixin Lai, Xuesheng Zhang, and Suiyao Chen. 2024. Adaptive ensembles of fine-tuned transformers for llm-generated text detection. *arXiv preprint arXiv:2403.13335*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yelena Mejova, Amy X Zhang, Nicholas Diakopoulos, and Carlos Castillo. 2014. Controversy and sentiment in online news. *arXiv preprint arXiv:1409.8152*.

Laura M Smith, Linhong Zhu, Kristina Lerman, and Zornitsa Kozareva. 2013. The role of social media in the discussion of controversial topics. In *2013 International Conference on Social Computing*, pages 236–243. IEEE.

Salvatore Vilella, Alfonso Semeraro, Daniela Paolotti, and Giancarlo Ruffo. 2021. The impact of disinformation on a controversial debate on social media. *arXiv preprint arXiv:2106.15968*.

6. Language Resource References

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. In *CoCo@ NIPS*.