Comment Filter: Steering the Debate in the News Comment Section by Promoting 'Good' User Comments

Johannes Filter
Hasso Plattner Institute, University Potsdam
Potsdam, Germany
Johannes.Filter@student.hpi.de

KEYWORDS

Online News Comments, User-generated Content, NLP, Machine Learning

1 INTRODUCTION

Online news outlets are drowning in the vast number of user comments. Trolling, toxic comments and out-of-topic discussions give the impression that humans are unable to debate in the feeling of pseudonymity. A lot of websites are closing their comment section, because they cannot afford to manually moderate it. The once emancipatory act of depriving journalists of their role as gatekeepers seems gone. But with the help of new natural-language processing (NLP) and machine learning techniques, there is hope to automatically analyze user comments. There already exists work to detect hate speech or abusive language [2, 15, 23] in user-generated content. In this work, we want to take a different perspective on the issue. Instead of eliminating 'bad' comments, we want to promote 'good' comments which are worth reading.

By doing so, we hope to steer the debate to be constructive and rational. We want to bring the discourse ethics[1] as proposed by Jürgen Habermas into practice. The principle of discourse ethics: a group of people with a rationale debate comes to common conclusion which manifest the morale. This is a start contract to Immanuel Kant's categorical imperative which focus on individuals. The theoretical concept of the discourse ethics can be set in practice in the comment section albeit in some variation. This is guiding principle of all of our work.

There is a long tradition of supporting journalistic work with digital technology. It started in the late 60s with computer-assisted reporting[2] and leading to current ideas about automatic reporting[3]. Right now, a big topic is supporting newspapers in managing their comments. There is e.g. the Coral Project[4] a cooperation among Mozilla, the New York Times and the Washington Post, that offer a range of open source tools. In the following section, related scientific work is examined.

[1]https://en.wikipedia.org/wiki/Discourse_ethics
[2]https://en.wikipedia.org/wiki/Computer-assisted_reporting
[3]https://en.wikipedia.org/wiki/Automated_journalism
[4]https://coralproject.net/

2 RELATED WORK

There is work done by Park et al. in the field of Human-Computer Interaction [17]. where they build an end-to-end system incorporating feature-based traditional machine learning. Recent work in the research area of comment analysis focuses on identifying high quality or constructive comments on e.g. New York Times website [3, 8, 9]. Other research focused on identifying valuable discourses [13]. Earlier work analyzed user-generated content on online services [5, 11, 25]. So far, most search focus on assessing the quality of a comment without considering the article. The recent of work by Cheng et al. [1] considers the abstract of the news article as well as surrounding comments. Closely related is the work by Qin et al. [21] who automatically generate high-quality comments.

Outside of the computer science community, there exists qualitative analysis of comments that should guide our work. Loosen et al. [12] formulated several comment quality indicators after conducting several interviews with News professionals. Earlier work by Diakopoulos et al. [4] and Noci et al. [16] highlight the quality of comments.

3 DATA

In the field of machine learning research, data is as important as the method. Thus, we give an overview over potential data sources.

3.1 Labeled Comment Corpura

There is a lack of labeled data for news comments. For English, there is the SOCC corpus [10] that labeled the constructiveness for 1k comments. The Yahoo annotated news corpus [14] is labeled on a thread level with 10k comment. For German, there is the 'One Million Post' [24] corpus of over 11k labeled comments of 'Der Standard'. Qin et al. [21] labeld over 40k Chinese comments for quality.

The idea of what constitutes a 'good' comment is not clearly defined. It can still be said, that out of those corpora, high-quality or good comments can be derived.

3.2 Generate Data

Although there exists some labeled data, the amount it still relatively small. Since in general, more data improves the performance of your deep learning technique [6], we want a big amount of data. For this, we look for a different way generating labeled data. Similiar to Cheng et al. [1] we want to consider the upvotes of a comment as a proxy for its quality. But in contrast to their approach, we normalize the upvotes. They simply say that comments with under 10 upvotes are negative samples and the rest are positive samples. We apply a more fine-grained preprocessing. In the following are the basic steps.

(1) Remove non-root comments
(2) Remove articles with few comments
(3) Remove articles with few up-votes
(4) Rank comments by chronological order
(5) Only consider first N comments per article
(6) Calculate relative upvotes for each comment
(7) Classify the comments with the most upvotes as positive and the least as negative, leaving out the middle

So the main is that, really good and really bad comments are most likely good and respectively bad. But for those in the middle, it's not clear so we leave them out. Potentially data sources are The Guardian, the New York Times and Zeit Online because their comments have upvotes.

3.3 Manually Label Data

Because there is not a enough labeled data, we could think about labeling it ourselves.

4 MACHINE LEARNING METHODS

In order to select comments that are worth promoting, we use a two-step process. First we select 'good' comments. Then we further filter those comments and promote only a selection to avoid duplicates.

For the first part, we want to investigate several machine learning methods ranging from traditional feature-based machine learning to state-of-the-art deep learning. We will try out already tested features on comments such as average word and average sentence length [9]. But also use stylometric features as used by Potthast et al. [20]. For the deep learning part, there is new idea of going away from simple vector embeddings with ELMO [19], Ulmfit [7], and a Transformer [22]. The general idea of those approaches: Train a network on a large corpus of unlabeld data and then finetune for your specific problem on labeled data. Perone et al. [18] compare the performance to several tasks.

For the second part, we cluster those comment and present only one (most likely the first one) out of each cluster. We will use topic modelling to find the clusters [not sure which one]. Right now it is open how to rank the comments. We may need additional labeled data for it.

5 EVALUATION

In order to evaluate the quality of the final model(s), we conduct a user study with a quantitative and a qualitative part and only compare the performance against a golden truth dataset.

In the first part, participants are required to answer questions about their background and online news usage. In the second part, people get presented two different ranking of comments on printed out paper. One half of the participants get the ordinary ranking first and the other our approach. This part is accompanied by a semi-structured interview to elicit information. To analyse the qualitative data, we encode the responses with thematic analysis.

In addition, we will manually label a small portion of comments as good and bad. Then, we can verify if our model.

6 SCOPE [IN INTRODUCTION?]

Developing an end-to-end system is out of scope for this work. The concrete implementation.

REFERENCES

[1] D. Chen, S. Ma, P. Yang, and X. Sun. 2018. Identifying High-Quality Chinese News Comments Based on Multi-Target Text Matching Model. *ArXiv e-prints* (Aug. 2018). arXiv:cs.CL/1808.07191
[2] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM '17)*. 512–515.
[3] Nicholas Diakopoulos. 2015. Picking the NYT Picks: Editorial criteria and automation in the curation of online news comments. (2015).
[4] Nicholas Diakopoulos and Mor Naaman. 2011. Towards Quality Discourse in Online News Comments. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work (CSCW '11)*. ACM, New York, NY, USA, 133–142. https://doi.org/10.1145/1958824.1958844
[5] Vicenç Gómez, Andreas Kaltenbrunner, and Vicente López. 2008. Statistical Analysis of the Social Network and Discussion Threads in Slashdot. In *Proceedings of the 17th International Conference on World Wide Web (WWW '08)*. ACM, New York, NY, USA, 645–654. https://doi.org/10.1145/1367497.1367585
[6] Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems* 24, 2 (2009), 8–12.
[7] Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 328–339.
[8] Varada Kolhatkar and Maite Taboada. 2017. Constructive language in news comments. In *Proceedings of the First Workshop on Abusive Language Online*. 11–17.
[9] Varada Kolhatkar and Maite Taboada. 2017. Using New York Times Picks to Identify Constructive Comments. Association for Computational Linguistics, 100–105. https://doi.org/10.18653/v1/W17-4218
[10] Varada Kolhatkar, Hanhan Wu, Luca Cavasso, Emilie Francis, Kavan Shukla, and Maite Taboada. 2018. The SFU Opinion and Comments Corpus: A Corpus for the Analysis of Online News Comments. (2018).
[11] Cliff Lampe and Paul Resnick. 2004. Slash(Dot) and Burn: Distributed Moderation in a Large Online Conversation Space. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. ACM, New York, NY, USA, 543–550. https://doi.org/10.1145/985692.985761
[12] Wiebke Loosen, Marlo H ring, Zijad Kurtanovi, Lisa Merten, Julius Reimer, Lies van Roessel, and Walid Maalej. 2017. Making sense of user comments: Identifying journalists requirements for a comment analysis framework. *Studies in Communication | Media* 6, 4 (2017), 333–364. https://doi.org/10.5771/2192-4007-2017-4-333
[13] Courtney Napoles, Aasish Pappu, and Joel R Tetreault. 2017. Automatically Identifying Good Conversations Online (Yes, They Do Exist!).
[14] Courtney Napoles, Joel Tetreault, Aasish Pappu, Enrica Rosato, and Brian Provenzale. 2017. Finding good conversations online: The Yahoo News annotated comments corpus. In *Proceedings of The 11th Linguistic Annotation Workshop*. 13–23.
[15] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection

in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web (WWW '16)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 145–153. https://doi.org/10.1145/2872427.2883062

[16] Javier Díaz Noci, David Domingo, Pere Masip, JL Micó, and C Ruiz. 2012. Comments in news, democracy booster or journalistic nightmare: Assessing the quality and dynamics of citizen debates in Catalan online newspapers. In *International Symposium on Online Journalism*, Vol. 2. 46–64.

[17] Deokgun Park, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. 2016. Supporting Comment Moderators in Identifying High Quality Online News Comments. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 1114–1125. https://doi.org/10.1145/2858036.2858389

[18] C. S. Perone, R. Silveira, and T. S. Paula. 2018. Evaluation of sentence embeddings in downstream and linguistic probing tasks. *ArXiv e-prints* (June 2018). arXiv:cs.CL/1806.06259

[19] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* (2018).

[20] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein. 2017. A Stylometric Inquiry into Hyperpartisan and Fake News. *ArXiv e-prints* (Feb. 2017). arXiv:cs.CL/1702.05638

[21] L. Qin, L. Liu, V. Bi, Y. Wang, X. Liu, Z. Hu, H. Zhao, and S. Shi. 2018. Automatic Article Commenting: the Task and Dataset. *ArXiv e-prints* (May 2018). arXiv:cs.CL/1805.03668

[22] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. (2018).

[23] Julian Risch and Ralf Krestel. 2018. Delete or not Delete? Semi-Automatic Comment Moderation for the Newsroom. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. 166–176.

[24] Dietmar Schabus, Marcin Skowron, and Martin Trapp. 2017. One Million Posts: A Data Set of German Online Discussions. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. ACM, New York, NY, USA, 1241–1244. https://doi.org/10.1145/3077136.3080711

[25] Gabor Szabo and Bernardo A. Huberman. 2010. Predicting the Popularity of Online Content. *Commun. ACM* 53, 8 (Aug. 2010), 80–88. https://doi.org/10.1145/1787234.1787254