University of Helsinki
Faculty of Arts
Department of Digital Humanities

Lecture: Computational Literacy

Date of Submission: 21.12.2022

Peer Review:
A Quantitative Analysis of Criticism

Name: Annika Grützner-Zahn
Mail: annika.grutzner-zahn@helsinki.fi

1 Introduction

Scientific work has to pass an evaluation conducted by other scholars from the field to be accepted by a publishing venue or by conference organisers. This practise of scientific evaluation through other peers started in the 18th century , but was established as a standard in the second half of the 20th century and is now conducted in all areas of research. Peer reviews were often criticised as subjective and in small research areas containing conflicts of interest (Benos et al., 2007; Lee et al., 2013) . Those critics led to many analysis and evaluations of peer reviews and their value. But apart from the research evaluation process itself, peer reviews contain also valuable information about biases, problems and mistakes in scientific work detected by the reviewers. Until now, the content of the peer reviews was not systematically analysed, although it could reveal important insights for especially young scholars about which mistakes are common and to avoid and what aspects are highlighted, if they are done in a good way. The positively highlighted aspects could also reveal important insights about what in valued high within a research community and whether this may contain biases with regard to specific kinds of scientific publications. In this work, the following research questions about the content of peer reviews will be addressed:

1. Which aspects of scientific works are repeatedly criticised in peer reviews?
2. Which aspects are often highlighted, if they are considered as extraordinary?

Given that research communities differ highly between different fields and subjects, the research questions will be limited to the subject of the dataset which contains reviews of publications about the field Natural Language Processing (NLP).

2 The Dataset: PeerRead

The dataset "PeerRead"[1] (Kang et al., 2018) was published 2018 on Github containing 14,782 papers and 10,770 reviews from conference proceedings in the research area NLP. A single publication contains the article itself, metadata about the article, the reviews and metadata about the reviews. The content was already processed from PDF files to JSON encoded text files which will facilitate the use of the data. A first overview of the data revealed some peculiarities in the content of the reviews, such as line breaks and LateX-commands which has to be cleaned. In addition, criteria and scores given are part of the reviews and also need to be cleaned.

The peer reviews were collected from four different conferences, NIPS 2013-2017 (9,152 reviews), ICLR 2017 (1,304 reviews), ACL 2017 (275 reviews) and CoNLL 2016 (39

[1] urlhttps://github.com/allenai/PeerRead

reviews). Except of the reviews from the NIPS conferences, the reviews are easily downloadable from Github. Because of licensing constraints, the files from the NIPS conference have to be downloaded using a python code snippet documented in the READ.ME of the Github folder. After the download of the data, it was recognised that the amount of files do not agree with the numbers provided by the authors (Kang et al., 2018). The review files imported from ACL 2017 were only 137, from CoNLL 2016 22 and from ICLR 2017 427. Each file contained several single reviews which were overall after the extraction of each review a list of 7584 single reviews from the conferences ACL 2017, CoNLL 2016 and ICLR 2017. In this first draft, only the data of those three conferences were used because of difficulties with the data collection from the NIPS conferences. Additionally, the benefit of implementing the pipeline with a smaller dataset is the reduced need of computing power. Even with the smaller dataset, the running and trying out of code took quite some time.

3 Methods of Data Processing

The Python Script used for the data processing can be found xx. The dataset was imported to a Jupyter Notebook[2] and filtered and cleaned using Python, the library Panda[3] and the RegEx module.[4] The dataset itself contained much more information than just the peer reviews and the information within the file was structured within a hierarchy of elements. To access the reviews, the hierarchy was solved using a panda dataframe and a following extraction of the relevant columns with the reviews. Regular expression were used to delete the content unrelevant for this project and to clean the reviews with regard to peculiarities described in Section 2 and special characters. A manual conducted check of the data revealed that it was not possible in the short amount of time to clean out every character unnecessary for the further processing steps. Given the planned next steps of extracting adjectives and single sentences, this is not mission critical to be able to conduct an analysis. To apply further processing steps to all reviews at a time, the dataframe was dissolved and one list containing all reviews were created. This list was saved and uploaded on Github.[5]

To be able to detect criticism and praise, evaluating adjectives were identified manually. For that, the reviews were tokenised and the adjectives identified using part-of-speech tagging. For this processing, the open Natural Language Processing library NLTK[6] was used. It was chosen because of the easy installation and application to the data. Before applying the library to the reviews, the reviews had to be converted to a string variable. The extracted adjectives were exported as a txt file and then categorised with regard to positive (pos) or negative (neg) evaluation, not evaluative (non) at all or false classified

[2]https://jupyter.org/
[3]https://pandas.pydata.org/
[4]https://www.w3schools.com/python/python_regex.asp
[5]https://github.com/Zahnanni/PeerReviews_NLP/blob/main/cleaned_data2.txt
[6]https://www.nltk.org/

(false). The difficulties of classifying adjectives perceived as being used often in positive and negative evaluative expressions lead to the inclusion of a fifth category: both. For example, the adjective "worth" could be used in a context like it would have been worth while to conduct an analysis deeper or in a context like this research was worth time and money. Another example is the adjective "difficult" which could be used to express difficulties understanding the content of the paper or to show that the topic was quite difficult but still well solved. Over 80% of the adjectives were classified and this list of adjectives and their attributions can be found on Github.[7]

Caused by time constraints, the project could be only conducted until here. Given that the data from the NIPS conference could not be included, it is planned to repeat the pipeline implemented so far with the data from the NIPS conference also. While classifying the main part of the adjectives, several issues with the data were recognised which will be solved in the next round of the project. The fllowing problems occured which can be cleaned and therefore reduce the effort of adjective classification:

- Missing space between dot and next word,
- missing space between word starting with an upper letter and another word,
- criterium still included: ISMETAREVIEW,
- numbers included through footnotes.

Additionally, before the adjectives will be extracted the next time, the upper letter will be lowered to reduce the amount of dublets. This time, only the dublets with the exact same writing were deleted. Another step will be the deletion of words with less than three letters, Acronyms and abbrevations. The inclusion of those steps will hopefully reduce the amount of noise in the list of adjectives significantly.

After the adjective classification, sentence segmentation will be applied to be able to extract the sentences containing the evaluative adjectives. Each sentence with a relevant adjective will be categorised into positive or negative feedback. The sentences extracted containing adjectives with the attribute "both" will be manually classified and added to the positive or negative list. Thus, in the end, there will be two lists of sentences containing the positive and negative criticism.

4 Analysis

It is planned to analyse the data based on a term extraction conducted using the tool annif.[8] The extracted terms shall reflect the most important terms used in connection with the adjectives. Afterwards, the terms can be statistically analysed to find the criticism mentioned several times for the positive and negative feedback.

[7] https://github.com/Zahnanni/PeerReviews_NLP/blob/main/adjectives.txt
[8] https://annif.org/

5 Results

Given that the analysis was not conducted yet, no results can be presented.

6 Discussion

Given the different research and publication habits, the dataset is not representative for scientific publications from other research fields. Additionally, the publications reviewed are all conference proceedings which may have an influence on the content because of the specific length required by the conference organisers. The representation of the conferences are unevenly distributed which may also cause an influence which can be hardly assessed without detailed information about the review process. The conference represented with the highest amounts of peer reviews also only published the peer reviews about accepted publications. Therefore, the dataset is biased with regard to accepted papers, meaning that the reviews may not contain the criticism leading to the rejection of a paper.

Biases which will be inserted through the processing is for one the exclusion of the feedback not containing an adjective, e.g. "In the same vein, pixel/feature control seems to have the most impact [...], I think it would have been worth looking at this, either in isolation or in more depth, measuring more than just performance on RL tasks". Moreover, the reduction to sentence-level pieces which will be extracted with the adjective may exclude important part of the feedback explained in more than one sentence.

The classification of adjectives into positive and negative is highly difficult without the context the adjective is used in. As explained in Section 3, the adjectives perceived as most likely to be used in both contexts were extra evaluated later in the pipeline. But the overall problem still persists all adjectives can be used together with a negation and therefore refer to the opposite of it's meaning. An analysis taking into account negation of adjectives could help at this point.

References

Benos, D. J., Bashari, E., Chaves, J. M., Gaggar, A., Kapoor, N., LaFrance, M., Mans, R., Mayhew, D., McGowan, S., Polter, A., Qadri, Y., Sarfare, S., Schultz, K., Splittgerber, R., Stephenson, J., Tower, C., Walton, R. G., & Zotov, A. (2007). The ups and downs of peer review. *Advances in Physiology Education*, *31*(2), 145–152. https://doi.org/10.1152/advan.00104.2006

Kang, D., Ammar, W., Dalvi, B., van Zuylen, M., Kohlmeier, S., Hovy, E., & Schwartz, R. (2018). A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications [Publisher: arXiv Version Number: 1]. https://doi.org/10.48550/ARXIV.1804.09635

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology*, *64*(1), 2–17. https://doi.org/10.1002/asi.22784