International Journal on Digital Libraries (2024) 25:493–504
https://doi.org/10.1007/s00799-023-00359-0

Towards automated meta-review generation via an NLP/ML pipeline in different stages of the scholarly peer review process

Asheesh Kumar[1] · Tirthankar Ghosal[2] · Saprativa Bhattacharjee[3] · Asif Ekbal[1]

Received: 25 November 2022 / Revised: 1 March 2023 / Accepted: 16 March 2023 / Published online: 24 April 2023

Abstract

With the ever-increasing number of submissions in top-tier conferences and journals, finding good reviewers and meta-reviewers is becoming increasingly difficult. Writing a meta-review is not straightforward as it involves a series of sub-tasks, including making a decision on the paper based on the reviewer's recommendation and their confidence in the recommendation, mitigating disagreements among the reviewers, and other such similar tasks. In this work, we develop a novel approach to automatically generate meta-reviews that are decision-aware and which also take into account a set of relevant sub-tasks in the peer-review process. More specifically, we first predict the recommendation scores and confidence scores for the reviews, using which we then predict the decision on a particular manuscript. Finally, we utilize the decision signals for generating the meta-reviews using a transformer-based seq2seq architecture. Our proposed pipelined approach for automatic decision-aware meta-review generation achieves significant performance improvement over the standard summarization baselines as well as relevant prior works on this problem. We make our codes available at https://github.com/saprativa/seq-to-seq-decision-aware-mrg.

Keywords Meta-review generation · Peer-review · Decision-aware meta reviews · Decision prediction

1 Introduction

Peer reviews are central for research validation, where multiple experts review the paper independently and then provide their opinion in the form of reviews. Sometimes the reviewers are required to provide their 'recommendation score' to reflect their assessment of the work. They are also sometimes required to provide their 'confidence score' to exhibit how familiar is the related literature to the reviewer or how confident the reviewer is about their evaluation. Not only the review text but also additional signals like recommendation and confidence scores from multiple reviewers help the chairs/editors to get a better feel of the merit of the paper and assist them in reaching their decision on the acceptance or rejection of the article to the concerned venue. The chair/editor then writes a meta-review cumulating the reviewers' views while justifying the decision on the paper's fate, finally communicating the same to the authors.

With the growing number of scientific paper submissions to top-tier conferences, having AI interventions [1] to counter the information overload seems justified. An AI assistant to generate an initial meta-review draft would help the chair to craft a meaningful meta-review quickly. Here in our initial investigation, we set out to investigate if we can leverage all the available signals to a human meta-reviewer (e.g., review-text, reviewer's recommendation, reviewer's conviction, reviewer's confidence [2], final judgment [3], and others) to automatically write a decision-aware meta-review. We design a deep neural architecture pipeline that performs

✉ Tirthankar Ghosal
ghosal@ufal.mff.cuni.cz

Asheesh Kumar
aseesnathhh@gmail.com

Saprativa Bhattacharjee
saprativa.bhatt@gov.in

Asif Ekbal
asif@iitp.ac.in

[1] Department of Computer Science and Engineering, Indian Institute of Technology Patna, Bihta, Patna, Bihar 801106, India

[2] Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Charles University, Malostranské náměstí 25, Prague 118 00, Czech Republic

[3] Department of Information Technology, Government Polytechnic Daman, Varkund, Daman, Dadra & Nagar Haveli and Daman & Diu 396210, India

the relevant sub-tasks at different stages in the pipeline to generate a decision-aware meta-review finally. Our primary motivation in this work is to replicate the stages in a human peer-review process while assisting the chairs in making informed decisions and writing good quality meta-reviews.

Specifically, we present a decision-aware transformer-based multi-encoder deep neural architecture to generate a meta-review while also predicting the final acceptance decision, which is again fueled by predicting the reviewer's sentiment, recommendation, conviction/uncertainty [4], and confidence in the intermediate steps. The multi-encoder gives three separate representations of the three peer-reviews for further input to the decoder. We use the review text and the reviewer's sentiment in our pipeline to predict the recommendation score. Then we use the predicted recommendation score along with the uncertainty of the reviewer (which we predict via a separate model [5]) to predict the confidence score. For each paper, we use the predicted recommendation score, uncertainty score, confidence score, sentiment, and representations of the three reviews to arrive at the final decision. Finally, we use the decision to generate a decision-aware meta-review. We evaluate our generated meta-reviews, both qualitatively and quantitatively. Although we achieved encouraging results, we emphasize that the current investigation is in its initial phase. We would require further fine-tuning and a deeper probe to justify our findings. Nevertheless, our approach to meta-review generation is novel and encompasses almost all the human factors in the peer review process.

The rest of the paper is organised as follows: Relevant prior works are discussed in Sect. 2. The dataset is described in Sect. 3. Our proposed methodology is presented in Sect. 4 along with a description of the sub-tasks incorporated in the pipeline. The evaluation metrics, baselines and comparing systems are described in Sect. 5. Results and analysis are given in Sect. 6. Finally, the conclusion is drawn in Sect. 7.

2 Related work

Although the problem is ambitious and new, there are a handful of investigations in the recent literature. The most relevant one being the decision-aware meta-review generation [6]. Here the authors mapped the three reviews to a high level encoder representation and used the last hidden states to predict decision while using a decoder to automatically generate the meta-review. In MetaGen [7], the authors first generate the extractive draft and then use a fine-tuned UniLM [8] (Unified Language Model) for the final decision prediction and abstractive meta-review generation. In [9], the authors investigate the role of summarization models and how far are we from meta-review generation with those large pre-trained models. We attempt the similar task, but we go one step further to perform multiple relevant sub-tasks in various stages of the peer-review process to automatically generate the meta-review, simulating the human peer-review process to a greater extent.

We also discuss some relevant works (decision prediction in peer reviews) that can add further context to the problem. The PeerRead [10] dataset is the first publicly available peer-review resource that encouraged Natural Language Processing (NLP)/ Machine Learning (ML) research on peer review problems. The authors defined two novel NLP tasks, viz. decision and aspect-score prediction [11]. Another work on conference paper acceptance prediction [12] extracted features from the paper such as title length, number of references, number of tables and figures, and others, to predict the final decisions using machine learning algorithms. The authors of DeepSentiPeer [13] used three channels of information: paper, corresponding review, and the review polarity to predict the overall recommendation as well as final decision. There are a few other works on NLP/ML for peer review problems [14, 15] such as aspect extraction [16] and sentiment analysis, which are worthy to explore to understand the related NLP/ML investigations in this domain.

3 Dataset

Research in the peer review system has been limited because of data privacy, confidentiality, and a closed system; however, in the last few years new open review system where reviews and comments along with the decision are posted publicly. This new process of review system has led to the availability of the data for studying the process.

3.1 Data collection

We collect the required peer review data (reviews, meta-reviews, recommendations, and confidence score) from the OpenReview[1] platform along with the decision of acceptance/rejection in the top-tier ML conference ICLR for the years 2018, 2019, 2020, and 2021. Most of the papers in our dataset have got three reviews. After pre-processing and eliminating some unusable reviews/meta-reviews, we arrive at 7,072 instances of papers with associated peer review data for our experiments. We use 15% of the data as the test set (1060), 75% as the training set (5304), and the remaining 10% as the validation set (708). Our proposed model treats each review individually (does not concatenate), so for training, we create a permutation in ordering the three reviews to have a training set of 31,824 reviews. We provide the total number of reviews, meta-reviews, and length in Table 1.

[1] https://openreview.net/.

Table 1 Details of the reviews and meta-review in our dataset across ICLR editions

Year	# Data	Max length	Min length	Avg length
2018	2802/934	2557/458	23/8	∼ 372.73/29.53
2019	4239/1413	4540/839	14/7	∼ 403.32/41.29
2020	6390/2130	3970/810	15/5	∼ 408.55/37.46
2021	7785/2595	4110/1102	14/5	∼ 455.65/52.25

Length is in the number of words. Each value in the row corresponds to statistics for review/meta-review

Table 2 Paper distribution for decision prediction task

Decision	Train	Test	Validation
Accept	1969	221	273
Reject	3688	415	506

Table 2 shows the distribution of reviews across paper-categories (Accepted or Rejected).

3.2 Data pre-processing

For recommendation and confidence labels, we normalize the values on the Likert scale of 1 to 5 and remove the label category when the data is less than 0.01%(rare class). Recommendation score of 1 means a strong reject, and 5 means a strong accept. Similarly, a confidence score of 1 indicates that the reviewer's evaluation is an educated guess either because the paper is not in the reviewer's area or was complicated to understand. On the other hand, a confidence score of 5 indicates the reviewer is absolutely sure that their evaluation is correct and that they are very familiar with the relevant literature. From Fig. 1, we can see that 85% of labels for recommendation score belong to only two classes and the rest three classes combined account for only 15%, confidence score distribution is similar with 78% taken up by two of the classes, leaving 22% only for the rest.

Ideally, a meta-review should contain all the key/deciding aspects collated from the multiple reviews along with the final decision on the concerned manuscript. Thus, we exclude those manuscripts from our dataset for which the meta-review or review word token size is less than 10, as we think such short meta-reviews/reviews contribute negligibly (sometimes even negatively) to the learning process.

4 Methodology

Peer-review decision is the central component of a meta-review. The chair writes the meta-review once they have already decided on the paper's fate. Hence, meta-reviews are decision-aware. We briefly discuss the sub-tasks in our pipeline in the subsequent sections.

4.1 The various prediction sub-tasks

Here we describe the three prediction sub-tasks which we utilize later (directly and indirectly) for aiding the main task of generating meta-reviews.

Recommendation Score Prediction We take the reviews along with sentence-level sentiment encodings to predict recommendation scores. In Table 3, we show examples of review sentences and their corresponding sentiment encodings (via VADER [17]) along with the final recommendation made by the corresponding reviewers. We can see that reviewer's sentiment (positive/negative/neutral) has a direct correlation to the final recommendation score. Hence, our decision to incorporate sentiment encodings for recommendation score prediction. We fine-tune a transformer-based pre-trained Bidirectional Encoder Representation from Transformer (BERT) [18] model for the given task. The BERT is a bidirectional transformer that is pre-trained using a combination of masked language modeling objective and next sentence prediction on a large corpus comprising the Toronto Book

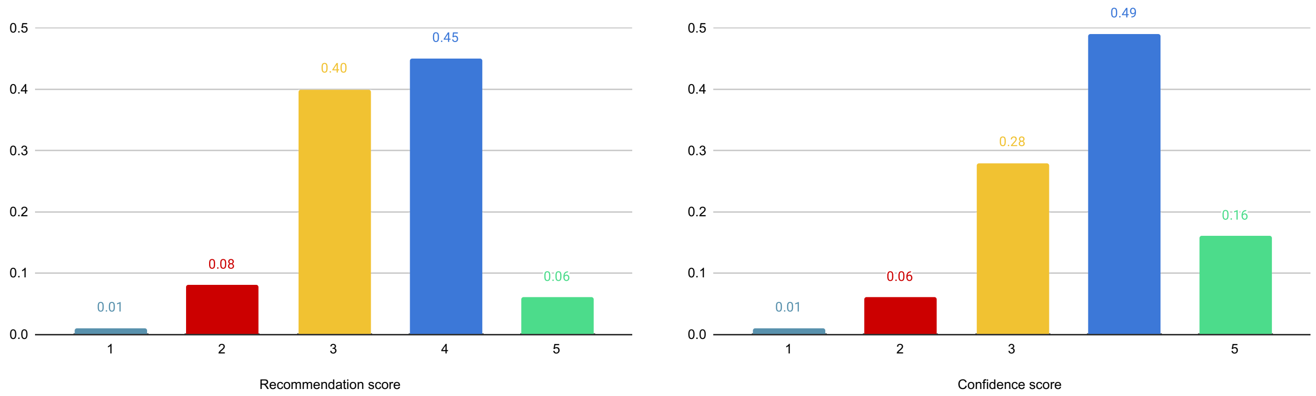


Fig. 1 Recommendation and confidence score normalized data distribution across labels

Table 3 Example of sentiment encoding from VADER for review sentences with corresponding recommendation scores

Review sentence	Sentiment [*‘comp’*, *‘neg’*, *‘neu’*, *‘pos’*]	Rec- Scores
Finally, the paper stops abruptly without any final discussion and/or conclusion	[−0.15, 0.14, 0.86, 0]	1
However, the results are not convincing and there is a crucial issue in the assumptions of the algorithm	[−0.31, 0.12, 0.88, 0]	1
[After author feedback] I would suggest that the authors revise the literature study and contributions to more accurately reflect prior work	[0, 0, 1, 0]	2
The paper sometimes uses L1 and sometimes L_1, it should be L_1 in all cases	[0, 0, 1, 0]	3
I find this paper both very interesting and important	[0.62, 0, 0.54, 0.46]	4
The paper is well written, the method is easy to implement, and the algorithm seems to have clear positive impact on the presented experiments	[0.88, 0, 0.64, 0.36]	4

Corpus and Wikipedia. BERT contextual representation of review augmented with Vader sentiment is given as input to feed-forward neural networks with ReLU, dropout and the batch norm as sub-layers with final layer after softmax doing multiclass classification to predict the scores.

Confidence Score Prediction For confidence score prediction, we take the review’s BERT representation, the predicted recommendation score, and the uncertainty score to predict the confidence score. For generating the uncertainty score, we use a pre-trained hedge-detection model [5] which uses XLNet [19] trained on BioScope Corpus [20] and SFU Review Corpus [21] to predict uncertain words. We use these predicted uncertain words and define an uncertainty score which is the ratio of the total number of uncertain word tokens in a review to the total words token in a review. We deem uncertainty/hedge cues from the reviewer as an important vertical to predict the reviewer’s confidence or conviction. We add the uncertainty score as a feature with the BERT contextual representation of the reviews and the recommendation scores to predict the confidence scores. We use these features as an input for our linear layer and then pass through dropout, batch norm and ReLU to a new linear layer and then to softmax for the final confidence score prediction. The model architecture used for recommendation score prediction and confidence score prediction is shown in Fig. 2.

Decision Prediction Finally, we build a model which takes three review representations, predicted recommendation score, predicted confidence score, and the specific sentence-level sentiment encodings of the review from VADER along with the predicted uncertainty score [5] as input to predict the

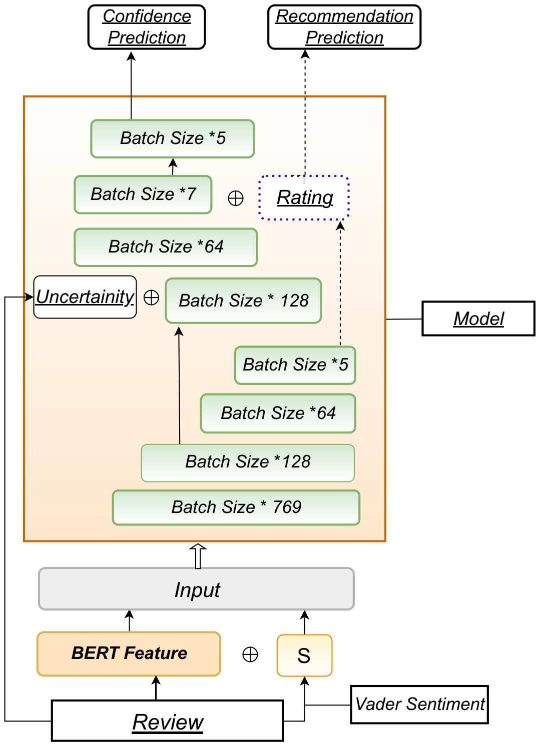


Fig. 2 Detailed architecture for recommendation score and confidence score prediction

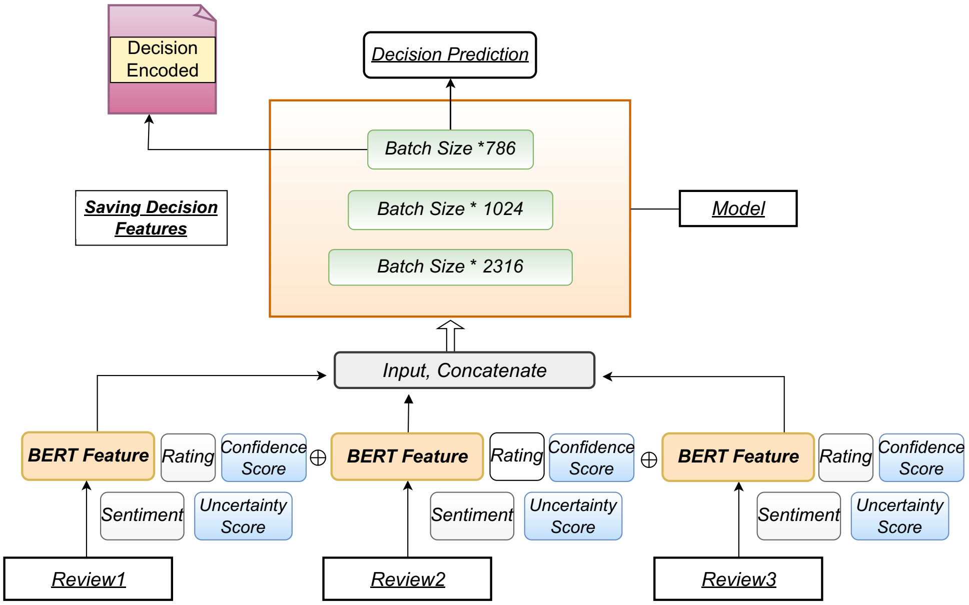


Fig. 3 Detailed architecture for decision prediction. Rating here refers to the predicted recommendation scores

final peer review decision on the paper. We present the model architecture for decision prediction in Fig. 3. The inputs are given to the linear layers from where they pass through ReLU, dropout, batch norm sub-layers and then to another linear layer for the final binary classification or the decision prediction.

4.2 Seq-to-seq meta-review generation: main task

Since our final problem is a generation one, we use transformer-based sequence-to-sequence encoder-decoder architecture for the generation task. As most papers have three reviews in our data, we use a transformer-based three encoders and a decoder model to automatically generate the meta-review. To make our multi-source transformer decision-aware, we use the former decision models' last encoding as input and pass it into decoder layers to provide the decision context (refer to Fig. 4).

Three encoders act as feature extractors that map the input vector to a high-level representation. With this representation, the decoder recursively predicts the sequence one at a time auto-regressively. The encoder consists of N layers of multi-head self-attention, feed-forward network, and residual connections. The decoder consists of M layers with sub-layers of multi-head self-attention, feed-forward, and extra cross-attention, also known as multi-head encoder-decoder attention. In a multi-source transformer, cross-attention with three past key-value pairs can be modeled in several ways [22]. We use a parallel strategy for our approach to produce a rich representation from the three encoders in the task. In addition, we choose to train a Byte-pair encoding tokenizer with the same special tokens as RoBERTa [23] and pick its vocab size to be 52,000.

5 Evaluation

In this section, we first describe the evaluation metrics that we chose for automatic evaluation of the model generated meta-reviews and then discuss about the selected baselines and comparing systems.

5.1 Evaluation metrics

To evaluate multi-class prediction (recommendation score and confidence score) and binary prediction (final decision) tasks, we use metrics such as accuracy, F1 score, and Root Mean Squared Error (RMSE) from scikit-learn.[2] While for the meta-review generation task, we use some popular automatic evaluation metrics which are used for evaluating text generation and summarization. Since a single metric does not give the best evaluation for a generated summary, we use ROUGE-1, ROUGE-2, ROUGE-3 [24], BERTScore [25], S3 [26] and BLEU [27] metrics. Below we describe the above mentioned evaluation metrics for meta-review generation:

ROUGE This is a widely adopted summarization evaluation metrics which stands for Recall-Oriented Understudy for Gisting Evaluation. ROUGE scores range from 0 to 1.0 means reference summary does not have any common n-gram unit with generated summary and 1 means all the n-gram units in reference summary has been captured by model generated summary. Thus, ROUGE-N measures unigram, bigram, trigram and higher order n-gram overlap between candidate and reference.

[2] https://scikit-learn.org/stable/.

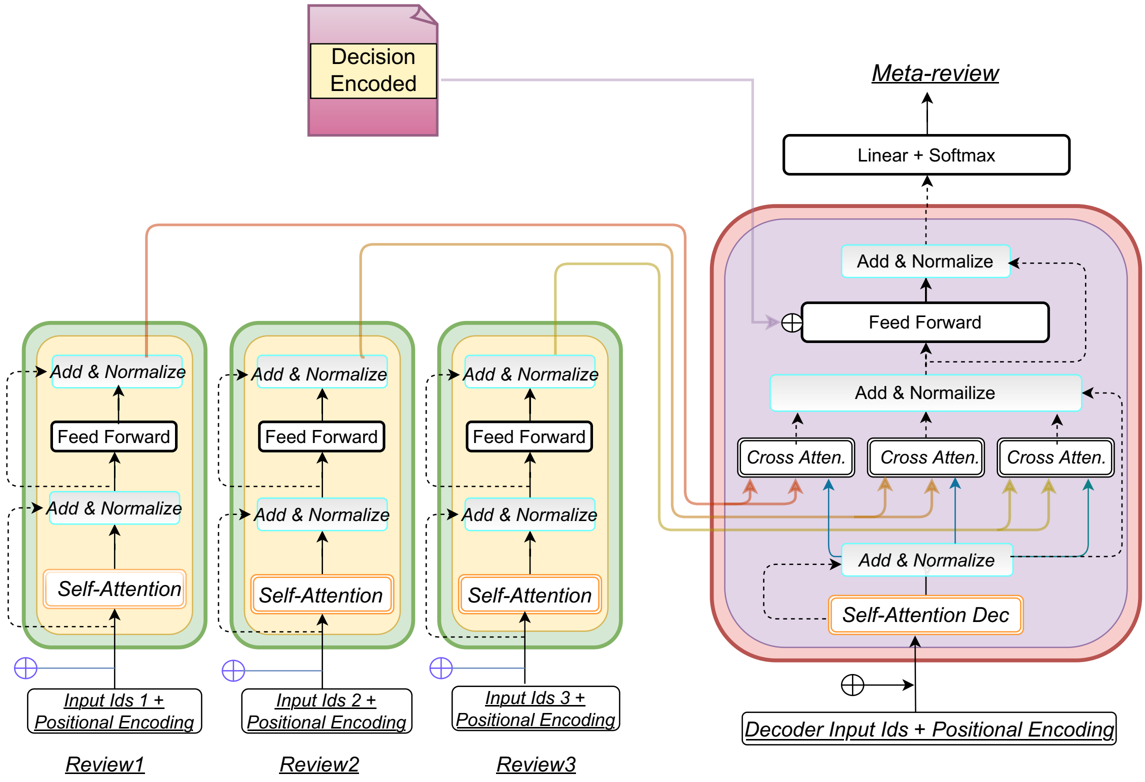


Fig. 4 Final model architecture of seq-to-seq decision-aware meta-review generation leveraging the encoded decision from Fig. 3

ROUGE-N recall between a system generated summary and a reference summary computes how much of the information contained in the reference summary is being captured by the system generated summary. It is calculated as follows:

$$\text{ROUGE-N}_{\text{recall}} = \frac{\text{\#matching n-grams between cand and ref}}{\text{\#n-grams in ref}} \tag{1}$$

On the other hand, ROUGE-N precision computes how much of the system generated candidate summary actually overlaps with the reference summary and is calculated as follows:

$$\text{ROUGE-N}_{\text{precision}} = \frac{\text{\#matching n-grams between cand and ref}}{\text{\#n-grams in cand}} \tag{2}$$

S3 This automatic scoring metrics creates a model trained on human judgment datasets from TAC conferences. It uses existing automatic metrics as features such as ROUGE, JS-divergence, and ROUGE-WE and predicts the score. The regression model learns the combination exhibiting the best correlation with human judgments. For experiments they have used Pyramid and the Responsiveness annotations. Models are trained and tested in leave one out cross validation ensuring proper testing of the approach for manual evaluation involving human in the process of scoring a reference summary with different scheme.

- **Responsiveness:** Human annotators score summaries on a LIKERT scale ranging from 1 to 5.
- **Pyramid:** Summarization Content Units (SCUs) are identified by noting/annotating information that are used for comparison of information in summaries. SCUs are variable in length but are not bigger than sentential clause, they emerge from annotation of a corpus of summaries for the same input. SCUs that appear in more manual summaries will get greater weights, so a pyramid will be formed after SCU annotation of manual summaries. The SCUs in peer summary are then compared against an existing pyramid to evaluate how much information agrees between the peer summary and manual summary. A key feature of a pyramid is that it quantitatively represents agreement among the human summaries.

BERTScore Evaluates generated text with pre-trained BERT contextual embeddings. BERTScore computes the similarity of candidate and reference summaries as a sum of cosine similarities between their token embeddings. Contextual embeddings gives different vector representations for the same words in different sentences, depending on the surrounding words, which form the context of the target word.

Given a reference sentence tokenized to k tokens $x = (x1, x2 \ldots, xk)$ and a candidate sentence tokenized to l tokens $\hat{x} = (\hat{x}1, \ldots, \hat{x}l)$ where each token is represented by contextual embeddings and calculates matching using cosine similarity. BERTScore computes recall score by matching token in reference sentence x to token in candidate sentence $\hat{x}$ and precision by matching the token in candidate sentence $\hat{x}$ to token in reference sentence x by using cosine similarity. F1 score is calculated by combining precision and recall. Uses greedy approach to match each token to the most similar token in the other sentence to maximize the similarity score.

For a reference x and candidate $\hat{x}$, the recall, precision, and F1 scores are:

$$R_{\text{BERTScore}} = \frac{1}{\text{mod}(x)} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^T \hat{x}_j \tag{3}$$

$$P_{\text{BERTScore}} = \frac{1}{\text{mod}(\hat{x})} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^T \hat{x}_j \tag{4}$$

$$F1_{\text{BERTScore}} = 2 * \frac{P_{\text{BERTScore}} * R_{\text{BERTScore}}}{P_{\text{BERTScore}} + R_{\text{BERTScore}}} \tag{5}$$

BLEU It is a widely used metrics in machine translation and it stands for Bilingual Evaluation Understudy. It is a precision-oriented metrics that calculates n-gram overlap between candidate and reference summary as follows:

$$\text{precision}_n = \frac{\sum\limits_{c \in \text{candidates}} \sum\limits_{n-\text{gram} \in c} \text{Count}_{\text{clip}}(\text{n-gram})}{\sum\limits_{c \in candidates} \sum\limits_{n-gram \in c} \text{Count}(\text{n-gram})} \tag{6}$$

$$\text{Count}_{\text{clip}}(\text{n-gram}) = \min(\text{matched n-gram count}, \max_{r \in \text{refs}}(\text{n-gram count in r})) \tag{7}$$

Using brevity penalty to penalize score with respect to the length of candidate summary. Brevity Penalty is multiplied to the so far score, that penalizes sentences that are shorter than any of our reference summary. This implies that if our candidate summary is as long as reference summary we multiply by 1.

$$\text{BP} = \begin{cases} 1, & \text{if } c > r \\ e^{1-(\frac{r}{c})}, & \text{if } c \leq r \end{cases} \tag{8}$$

$$\text{BLEU-N} = \text{BP} * \exp\left(\sum_{n=1}^{N} W_n \log\ \text{precision}_n\right) \tag{9}$$

where N is the number of n-grams which is by default 4 and W_n are the weights of the different n-gram precisions.

5.2 Baselines and comparing systems

Our initial experiments include PEGASUS, a BART-based summarizer, which we treat as the baseline for comparison, and two variants of our proposed model. We also use a pre-trained decision-aware MRG model and predict our test data.

We keep the learning rate for experiments as 5e-05, the number of beams for $beamsearch = 4$, $loss = crossentropy$, and $optimizer = Adam$. We train different models for 100 epochs with learning rate scheduler=linear and choose the best variant in terms of validation loss.

PEGASUS [28] PEGASUS is an abstractive summarization algorithm which uses self-supervised objective Gap Sentences Generation (GSG) to train a transformer-based encoder-decoder model. In PEGASUS, important sentences are removed/masked from an input document and are generated together as one output sequence from the remaining sentences, similar to an extractive summary. The best PEGASUS model is evaluated on 12 downstream summarization tasks spanning news, science, stories, instructions, emails, patents, and legislative bills. Experiments demonstrate that it achieves state-of-the-art performance on all 12 downstream datasets measured by ROUGE scores.

BART [29] BART uses a standard transformer-based seq2seq architecture with a bidirectional encoder and a unidirectional decoder. The pre-training task involves randomly shuffling the order of the original sentences and a novel in-filling scheme, where text spans are replaced with a single mask token. BART is particularly effective when fine-tuned for text generation and works well for comprehension tasks. We use the Hugging Face implementation of 12 encoder layers and 12 decoder layers with pre-trained weights[3] and fine-tune them on our dataset to generate the meta-review.

Simple Meta-Review Generator (S-MRG) This is a simple transformer-based architecture with only three encoders, each with two encoder layers to map inputs to a high-level representation and a decoder of two decoder layers with softmax normalization applied on the last hidden state in decoder for generating the sequence probability distribution over whole target vocabulary recursively, one at a time auto-regressively.

Decision-Aware MRG (MRG Decision) [6] MRG Decision predicts the decision from encoders' hidden states and carries the decision vector encoded from the encoder-hidden state output to the decoder layer, to provide the context to the

[3] https://huggingface.co/transformers/model_doc/bart.html.

Table 4 Model scores for automatic evaluation metrics

Model	ROUGE-1 (R/P)	ROUGE-2 (R/P)	ROUGE-3 (R/P)	S3 (pyr/resp)	BERTScore (f1)	BLEU
Pegasus [28]	0.19/0.38	0.04/0.08	0.01/0.01	0.10/0.31	0.54	2.14
BART [29]	0.32/0.41	0.08/0.10	0.02/0.03	0.24/ 0.39	0.57	2.85
S-MRG	0.27/0.35	0.05/0.06	0.01/0.01	0.16/0.33	0.55	1.50
MRG Decision [6]	0.30/0.36	0.06/0.07	0.01/0.01	0.19/0.35	0.55	1.75
Proposed approach: $S2S_{MRG}$	0.31/0.43	0.06/0.09	0.01/0.02	0.20/0.35	0.56	2.90

The output is the average of all the scores in the test set. R and P refers to recall and precision

Table 5 Results with respect to F1 score and overall accuracy for decision prediction, where S → sentiment and H → uncertainty score

Model	Accept	Reject	Accuracy
Review Text	0.43	0.79	0.69
Review+Recommendation+Confidence	0.75	0.83	0.82
DeepsentiPeer [13]	0.71	0.74	0.73
MRG Decision [6]	0.29	0.75	0.63
Review+Recommendation+Confidence+S+H	**0.76**	**0.88**	**0.84**

Bold values indicate better F1 scores (for Accept and Reject classes) as well as better accuracy as compared to the preceding models

generator module. The decoder's last hidden state after the softmax layer predicts the sequence recursively.

Proposed Approach/Model: Seq-to-Seq Decision Aware Meta-Review Generation ($S2S_{MRG}$) We improve the decision prediction model by using various input features as we notice that MRG Decision lacks in decision making *(accuracy of 63 %)*. In Fig. 4 our model $S2S_{MRG}$ uses the decision encoded vectors in all decoder layers where vectors are concatenated before the feed-forward sub-layer to provide the context to the generator module.

Our proposed approach takes input from the decision-prediction module (hence *decision-aware* just as human chairs do) to generate the meta-reviews.

6 Results and analysis

We present the results of automatic evaluation of the model generated meta-reviews along with a comparison with the selected baselines as well as comparing systems in Table 4. In Table 5 we present the performance comparison of our decision prediction task with various combination of features and comparable systems.

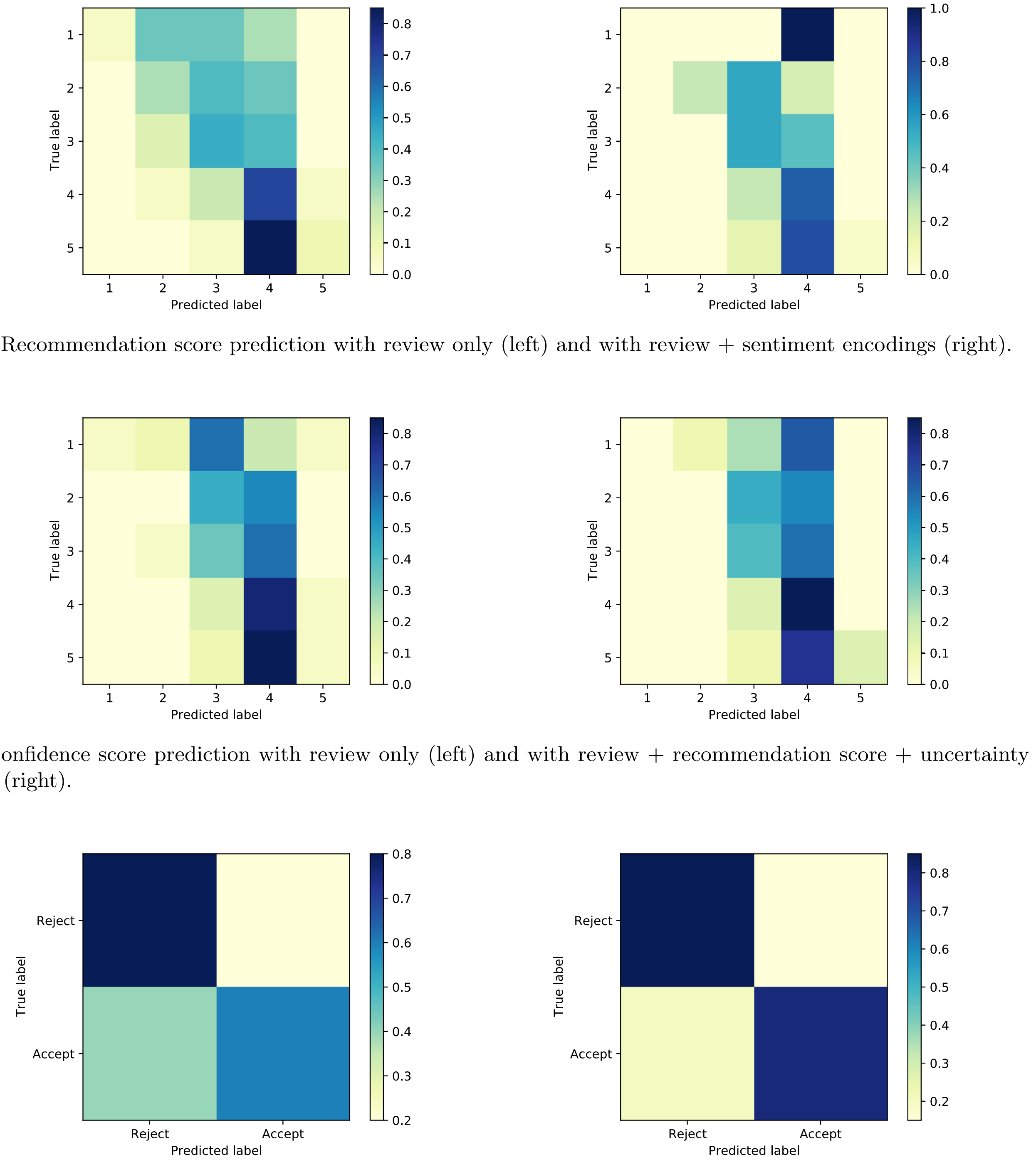
6.1 Quantitative and qualitative analysis

Our seq-to-seq meta-review generation model outperforms all the baseline models for ROUGE precision and BLEU scores. We achieve comparable results with the BART-based summarization model for all other scores. However, we argue that the evaluation is unfair as MRG and summarization are not the same tasks. We also evaluated the previous decision-aware model for meta-review generation MRG Decision [6] and found that our model outperforms it in terms of all quantitative metric scores.

The Root Mean Squared Error (*RMSE*) for the recommendation score prediction task when we use only review text comes to be 0.76. While when we use sentence-level sentiment encoding along with review text, our RMSE reduces to 0.75. For sentence-level sentiment encoding examples, refer to Table 3. For the confidence score prediction task, when we predict only using review text, we obtain an RMSE of 0.86. Further, when we incorporate recommendation and uncertainty scores along with the review text, the RMSE reduces to 0.82.

From Table 5, we can see that the final decision module also improves by 21% with respect to MRG Decision. In terms of the various feature combinations, our decision prediction model accuracy improves by 15% when we use several input features such as the recommendation, confidence, hedge score, sentiment encodings, and the text of the three reviews instead of simply using the review text.

The confusion matrices for the three sub-tasks are shown in Fig. 5. It is evident that when we incorporate sentiment encodings for recommendation score prediction and the predicted recommendation scores along with uncertainty scores for the confidence score predictions, the predictions get concentrated over the classes 3 and 4 (see the confusion matrices on the right side of top and middle rows). This is expected and is perfectly aligned with the data distribution in the dataset where more than 78% of the data belongs to these two classes as depicted in Fig. 2. Moreover, for the task of decision prediction, we observe that taking the predicted recommendation score, predicted confidence score, uncertainty score and sentiment encodings along with the review text into



(a) Recommendation score prediction with review only (left) and with review + sentiment encodings (right).

(b) Confidence score prediction with review only (left) and with review + recommendation score + uncertainty score (right).

(c) Decision prediction with review only (left) and with review + recommendation score + confidence score + sentiment encodings + uncertainty (right).

Fig. 5 Confusion matrices of the three prediction sub-tasks

consideration results in improved prediction accuracy for the “Accept" class, thus improving the overall accuracy.

Table 6 shows the MRG outputs of the different techniques. We use the pre-trained models for PEGASUS and BART from HuggingFace[4] but fine-tune on our review dataset. Our custom architectures with two different setups are entirely trained on our dataset. We find that although

[4] https://huggingface.co/.

Table 6 Ground truths and automatically generated meta-review for a given paper

Original Meta-Review:

https://openreview.net/forum?id=B1liraVYwr

This paper tackles neural response generation with Generative Adversarial Nets (GANs), and to address the training instability problem with GANs, it proposes a local distribution oriented objective. The new objective is combined with the original objective, and used as a hybrid loss for the adversarial training of response generation models, named as LocalGAN. Authors responded with concerns about reviewer 3's comments, and I agree with the authors explanation, so I am disregarding review 3, and am relying on my read through of the latest version of the paper. The other reviewers think the paper has good contributions, however they are not convinced about the clarity of the presentations and made many suggestions (even after the responses from the authors). I suggest a reject, as the paper should include a clear presentation of the approach and technical formulation (as also suggested by the reviewers).

PEGASUS [28]:

ICLR 2018 Conference Acceptance Decision. The reviewers have unanimously expressed strong concerns about the novelty of the paper.. Reject

BART [29]:

This paper proposes a new method for training a generative model that is robust to adversarial perturbations. The reviewers and AC note the critical limitation of novelty of the paper to meet the high standard of ICLR. AC thinks the proposed method has potential and is interesting, but decided that the authors need more works to publish.

S-MRG:

This paper proposes a novel application of generative adversarial networks to model neural network generation with arbitrary conditional autoencoders. While the reviewers initially some concerns regarding the motivation of the work, they unanimously agree that the paper was a quite ready for publication in its current form. In particular, the paper is hard to follow and understand the use of GANs, and the contributions is unclear.

MRG Decision [6]:

The paper proposes a novel method for improving generative properties of GAN training. The reviewers unanimously agree that this paper is not ready to be published, particularly being concerned about the unclear objective and potentially misleading claims of the paper. Multiple reviewers pointed out about incorrect claims and statements without theoretical or empirical justification. The reviewers also mention that the paper does not provide new insights about applicability of the method.

S2S$_{\mathbf{MRG}}$:

This paper proposes a method to train a neural network that uses the weights of a generative model, which can be used to generate the input. The method is evaluated on several datasets. The reviewers and AC agree that the paper is not well written. However, there are some concerns about the novelty of the proposed method and the experimental results are not convincing.

PEGASUS generated meta-review manifests sentences with polarity, the output is not detailed. The significant aspects of concern in the human-generated review are not prominent in the generated meta-review. The overall polarity and the decision do not match with the original meta-review. On the other hand, we observe that the output with BART, which is an extensive language model with 406 million parameters is detailed. Moreover, the generated meta-review also manifest polarity, and highlight merits and demerits. Our model S-MRG, does a reasonable job of capturing the polarity (see Table 6), and also the generated meta-review is in the third person. However, we notice that some irrelevant text from other papers' common primary keywords is present in the generated meta-review, which is eventually noise in the output.

Although the Decision-aware MRG [6] model writes the meta-review in the third person/as meta-reviewer in coherence with the existing peer reviews, but its decision prediction module has an accuracy of only 63%. Our proposed seq-to-seq decision-aware MRG model outputs are detailed and write the meta-review in the third person/as meta-reviewer in coherence with the existing peer reviews and brings out the merits and demerits of the paper. Generated meta-review also manifests polarity. The decision prediction module has a higher accuracy of 84%, which can be further improved by augmenting review-paper interaction as additional information channels to the model. We argue that the decision prediction module plays a key role in helping the model generate meta-reviews with the correct connotation as the meta-reviews are generally written by the chairs/editors only after a decision regarding the fate of the manuscript has already been made. Hence, higher the decision accuracy, higher the chance of generating a better meta-review.

6.2 Error analysis

We perform initial error analysis on our generated output. The automatically generated meta-review sometimes contains repeating texts. We also found that for a few papers, the ground truth decision is a reject. However, the generated meta-review by the model recommends accepting the paper. Sometimes meta-reviewers write from outside the context of the reviews or one-liners with a positive connotation (example: *The work brings little novelty compared to existing literature.*), but the final decision is negative. In such cases, our model fails, probably due to the lack of proper context.

We would look forward to doing a more in-depth analysis of our errors.

7 Conclusion

In this preliminary investigation, we propose a new technique for incorporating decision-awareness for automatic generation of meta-reviews from the peer reviews of manuscripts. We do this by taking into account the various sub-tasks which form an integral part of a human peer review process. Specifically, we first predict the recommendation scores based on the review texts and their sentiments. We then use the predicted recommendation scores along with uncertainty scores to predict the confidence scores of the respective reviews. Next, we use these scores and other features to predict the final decision on the manuscript. With the incorporation of these intermediate sub-tasks, we obtain an improvement of 21% in the decision prediction task, which is crucial to meta-review generation. Finally, we use the predicted decisions to generate the meta-reviews. Our proposed approach outperforms the earlier works and performs comparably with BART, which is a large complex neural architecture with 12 encoders and 12 decoders. However, we agree that only text summarization does not simulate a complex task such as automatically writing meta-reviews. As our immediate next step, we would like to deeply investigate fine-tuning of the specific sub-tasks, use the final-layer representations of the sub-tasks instead of the predictions, and perform a sensitivity analysis of each sub-task on the main task. Additionally, we would like to incorporate more fine-grained decisions such as strong/weak accept/reject or minor/major revisions instead of binary decisions. Finally, we would also like to explore a multi-tasking framework for meta-review generation in the future.

Acknowledgements Tirthankar Ghosal is funded by Cactus Communications, India (Award # CAC-2021-01) to carry out this research. Asif Ekbal receives the Visvesvaraya Young Faculty Award, thanks to the Digital India Corporation, Ministry of Electronics and Information Technology, Government of India for funding him in this research.

References

1. Ghosal, T.: Exploring the implications of artificial intelligence in various aspects of scholarly peer review. Bull. IEEE Tech. Comm. Digit. Libr. 15 (2019)
2. Bharti, P.K., Ghosal, T., Agrawal, M., Ekbal, A.: How confident was your reviewer? estimating reviewer confidence from peer review texts. In: Uchida, S., Barney, E., Eglin, V. (eds.) Document Analysis Systems, pp. 126–139. Springer, Cham (2022)
3. Bharti, P.K., Ranjan, S., Ghosal, T., Agrawal, M., Ekbal, A.: Peerassist: Leveraging on paper-review interactions to predict peer review decisions. In: Towards Open and Trustworthy Digital Societies: 23rd International Conference on Asia-Pacific Digital Libraries. ICADL 2021, Virtual Event, December 1–3, 2021, Proceedings, pp. 421–435. Springer, Berlin (2021)
4. Ghosal, T., Varanasi, K.K., Kordoni, V.: Hedgepeer: a dataset for uncertainty detection in peer reviews. In: Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries. JCDL '22. Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3529372.3533300
5. Britto, B.K., Khandelwal, A.: Resolving the scope of speculation and negation using transformer-based architectures. CoRR arXiv:2001.02885 (2020)
6. Kumar, A., Ghosal, T., Ekbal, A.: A deep neural architecture for decision-aware meta-review generation. In: 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 222–225 (2021). IEEE
7. Bhatia, C., Pradhan, T., Pal, S.: Metagen: An academic meta-review generation system. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1653–1656 (2020)
8. Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., Hon, H.: Unified language model pre-training for natural language understanding and generation. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada, pp. 13042–13054 (2019)
9. Bharti, P.K., Kumar, A., Ghosal, T., Agrawal, M., Ekbal, A.: Can a machine generate a meta-review? how far are we? In: Text, Speech, and Dialogue (TSD). Springer, Cham (2022)
10. Kang, D., Ammar, W., Dalvi, B., van Zuylen, M., Kohlmeier, S., Hovy, E.H., Schwartz, R.: A dataset of peer reviews (peerread): Collection, insights and NLP applications. In: Walker, M.A., Ji, H., Stent, A. (eds.) Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, June 1–6, 2018, Volume 1 (Long Papers), pp. 1647–1661. Association for Computational Linguistics, New Orleans, Louisiana, USA (2018). https://doi.org/10.18653/v1/n18-1149
11. Kumar, S., Ghosal, T., Bharti, P.K., Ekbal, A.: Sharing is caring! joint multitask learning helps aspect-category extraction and sentiment detection in scientific peer reviews. In: 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 270–273 (2021). https://doi.org/10.1109/JCDL52503.2021.00081
12. Joshi, D.J., Kulkarni, A., Pande, R., Kulkarni, I., Patil, S., Saini, N.: Conference paper acceptance prediction: Using machine learning. Machine Learning and Information Processing: Proceedings of ICMLIP 2020 1311, 143 (2021)
13. Ghosal, T., Verma, R., Ekbal, A., Bhattacharyya, P.: Deepsentipeer: Harnessing sentiment in review texts to recommend peer review decisions. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1120–1130 (2019)
14. Ghosal, T., Kumar, S., Bharti, P.K., Ekbal, A.: Peer review analyze: a novel benchmark resource for computational analysis of peer reviews. PLoS ONE **17**(1), 0259238 (2022). https://doi.org/10.1371/journal.pone.0259238
15. Bharti, P.K., Ghosal, T., Agrawal, M., Ekbal, A.: Betterpr: A dataset for estimating the constructiveness of peer review comments. In: Linking Theory and Practice of Digital Libraries (TPDL). Springer, Cham (2022)
16. Verma, R., Shinde, K., Arora, H., Ghosal, T.: Attend to your review: A deep neural network to extract aspects from peer reviews. In: Mantoro, T., Lee, M., Ayu, M.A., Wong, K.W., Hidayanto, A.N. (eds.) Neural Information Processing, pp. 761–768. Springer, Cham (2021)
17. Hutto, C.J., Gilbert, E.: VADER: A parsimonious rule-based model for sentiment analysis of social media text. In: Adar, E., Resnick, P., Choudhury, M.D., Hogan, B., Oh, A. (eds.) Proceedings of the

Eighth International Conference on Weblogs and Social Media, ICWSM 2014, June 1–4, 2014. The AAAI Press, Ann Arbor, Michigan, USA, (2014)

18. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, June 2–7, 2019, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, MN, USA (2019). https://doi.org/10.18653/v1/n19-1423
19. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
20. Szarvas, G., Vincze, V., Farkas, R., Csirik, J.: The bioscope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, pp. 38–45 (2008)
21. Konstantinova, N., de Sousa, S.C.M., Díaz, N.P.C., López, M.J.M., Taboada, M., Mitkov, R.: A review corpus annotated for negation, speculation and their scope. In: Calzolari, N., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, May 23–25, 2012, pp. 3190–3195. European Language Resources Association (ELRA), Istanbul, Turkey (2012)
22. Libovický, J., Helcl, J., Marecek, D.: Input combination strategies for multi-source transformer decoder. In: Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Jimeno-Yepes, A., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M.L., Post, M., Specia, L., Turchi, M., Verspoor, K. (eds.) Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, October 31 - November 1, 2018, pp. 253–260. Association for Computational Linguistics, Belgium, Brussels (2018). https://doi.org/10.18653/v1/w18-6326
23. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR arXiv:1907.11692 (2019)
24. Lin, C.-Y.: ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (2004). https://aclanthology.org/W04-1013
25. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with BERT. In: 8th International Conference on Learning Representations, ICLR 2020, April 26-30, 2020. OpenReview.net, Addis Ababa, Ethiopia (2020)
26. Peyrard, M., Botschen, T., Gurevych, I.: Learning to score system summaries for better content selection evaluation. In: Proceedings of the Workshop on New Frontiers in Summarization, pp. 74–84 (2017)
27. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)
28. Zhang, J., Zhao, Y., Saleh, M., Liu, P.: Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In: International Conference on Machine Learning, pp. 11328–11339 (2020). PMLR
29. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, July 5–10, 2020, pp. 7871–7880. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.703

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.