A Method for Automatically Estimating the Informativeness of Peer Review

Prabhat Kumar Bharti[1], **Tirthankar Ghosal**[2], **Mayank Agarwal**[1], **Asif Ekbal**[1]
[1]Department of Computer Science and Engineering, Indian Institute of Technology Patna, India
[2] UFAL, MFF, Charles University, Czech Republic
prabhat_1921cs32@iitp.ac.in, ghosal@ufal.mff.cuni.cz
mayank265@iitp.ac.in, asif@iitp.ac.in

Abstract

Peer reviews are intended to give authors constructive and informative feedback. It is expected that the reviewers will make constructive suggestions over certain aspects, e.g., novelty, clarity, empirical and theoretical soundness, etc., and sections, e.g., problem definition/idea, datasets, methodology, experiments, results, etc., of the paper in a detailed manner. With this objective, we analyze the reviewer's attitude toward the work. Aspects of the review are essential to determine how much weight the editor/chair should place on the review in making a decision. In this paper, we used a publicly available Peer Review Analyze dataset of peer review texts manually annotated at the sentence level (13.22 k sentences) across two layers: Paper Section Correspondence and Paper Aspect Category. We transform these categorical annotations to derive an informativeness score of the review based on the review's coverage across section correspondence, aspects of the paper, and reviewer-centric uncertainty associated with the review. We hope that our proposed methods, which are motivated towards automatically estimating the quality of peer reviews in the form of informativeness scores, will give editors an additional layer of confidence for the automatic judgment of review quality. We make our codes available at https://github.com/PrabhatkrBharti/informativeness.git.

1 Introduction

The peer review process is the central mechanism for validating scientific research (Siler et al., 2015). A good review typically provides feedback on one or more sections and aspects while reviewing the manuscript/paper [1], rather than just one section, say the Introduction (Kühne et al., 2010). Therefore, reviews covering more sections and aspects are more likely helpful to the author. Furthermore, the more sections and aspects the review covers, the higher the expected coverage score. It may give the author a confidence that the reviewer has read through and paid attention to the different sections and aspects in their submission. In addition, the reviewers are expected to provide constructive comments and suggestions regarding certain aspects and sections of the manuscript. To determining whether the reviewer was informative or constructive in their review and covered significant sections of the manuscript. It would be appropriate to mention the data from Peer Review Analyze (Ghosal et al., 2022a). They analyze and understand the reviewers' thrust over specific sections and aspects of the manuscript. We use those insights in our proposed method. This particular motivation led us to incorporate the general sections, and aspects of the paper defined by the Peer Review Analyze (Ghosal et al., 2022a) into this paper to calculate the informativeness score. We attempt here to generate an informativeness score for a given review directly by analyzing the review's coverage across section correspondence, aspects of the paper, and reviewer-centric uncertainty associated with the review.

We summarize the key contributions of this work as follows.

- We propose a seed idea for the automatic judgment of review quality.
- We introduce a novel method for measuring the informativeness score based on sections, aspects coverage, and reviewer-centric uncertainty encapsulated in the review.
- In addition, we establish statistical-driven baselines to evaluate Mean absolute error (MAE), Root Mean Square Error (RMSE) and coefficient of determination (R^2).

The novelty of our work lies in utilizing the Peer

[1]In this manuscript, manuscript/paper are used interchangeably.

Proceedings of the 19th International Conference on Natural Language Processing (ICON), pages 280 - 289
December 15-18, 2022 ©2022 Association for Computational Linguistics

Review Analyze dataset for measuring the informativeness score. Although we use the reviews of a premier machine learning conference (ICLR) as our dataset, our proposed method would represent a generic aspect of peer review in Science, Technology, Engineering and Mathematics (STEM). It will assist the editors in which review they should pay more attention to when crafting a meta-review. In addition, it may give the author confidence that if the review has high informativeness score, it means the reviewer has reviewed thoroughly their submission.

2 Related Work

In the Meta Science community and Peer Review Congress[2] (Brezis and Birukou, 2020), peer review quality has been a major research topic since 1989. There are a few relevant ones that we discuss in this article. The authors (Justice et al., 1998) studied a randomized control trial to see how masking author identity improves peer review quality. The study in(Jefferson et al., 2002) presented approaches for assessing the quality of editorial peer reviews. To assess peer reviews of manuscripts, the authors of (Van Rooyen et al., 1999) developed the Review Quality Instrument (RQI). In this paper, the authors (Shattell et al., 2010) examined the perspectives of authors and editors on the quality of peer review in three scholarly nursing journals. Peer review quality is evaluated in (Van Rooyen, 2001). A systematic review and meta-analysis on the impact of interventions to improve the quality of peer reviews of biomedical journals were conducted in (Bruce et al., 2016). In this paper, authors (Enserink, 2001) explored the dubious connection between the peer review and quality. Authors (D'Andrea and O'Dwyer, 2017) argued if the editors can save peer reviews from peer reviewers. (Rennie, 2016) advocates scientific guidelines for peer review. The purpose of this (Callaham et al., 1998) study was to evaluate the reliability of the editor's opinion subjective quality ratings of peer review of manuscripts. This paper provides an overview of how peer-review reports of scientific articles can be assessed by the authors (Sizo et al., 2019). For peer reviews, some relevant NLP/ML works are worth exploring from an NLP/ML perspective (Kumar et al., 2021; Ghosal et al., 2019; Ghosal, 2019; Kumar et al., 2022; Ghosal et al., 2022b; Bharti et al., 2022a,b, 2021; Gao et al., 2019). It should be noted, however, that none of these works attempted to determine the quality of peer reviews based on linguistic aspects. Here, the goal is to derive a justifiable informativeness score and then use those insights to investigate further, enabling editors to automatically identify the quality of peer reviews.

[2]https://peerreviewcongress.org/

3 Dataset

The dataset used in this study is from Peer Review Analyze (Ghosal et al., 2022a), which is publicly available. In Peer Review Analyze, peer review texts are manually annotated at the sentence level (13.22k sentences) across two layers: Paper Section Correspondence and Paper Aspect Category. The detailed dataset statistics are presented in Table 1, and the reader is referred to the original paper for further information.

3.1 Proposed Method

As we review the standard guidelines [3,4,5,6] for peer-reviewing in machine learning (ML) and natural language processing (NLP) conferences, we learn that the community expects a good review that covers more sections and aspects of the reviewed manuscript (Gregory and Denniss, 2019; Kühne et al., 2010). Having this motivation led us to develop a justifiable informativeness score which enables editors to automatically identify good reviews and isolate those that are less thorough. In our view, a good peer review should comment on key sections and highlight the reviewer's perspective while focusing on the essential aspects of the manuscript.

Peer Review Analyze dataset is used to generate an informativeness score based on the coverage of section correspondence, aspects of the paper, and the reviewer-centric uncertainty inherent in the review.

Paper Section Correspondence: The paper section correspondence identifies the section of the paper on which the review statement is commenting. E.g, Abstract (ABS), Introduction (INT), Related Works (RWK), Problem Definition/Idea (PDI), Data/Datasets (DAT), Methodology (MET), Experiments (EXP), Results (RES), Tables & Figures (TNF), Analysis (ANA), Future Work (FWK),

[3]https://iclr.cc/Conferences/2022/MetareviewGuide
[4]https://acl2020.org/reviewers/
[5]https://neurips.cc/Conferences/2022/ReviewerGuidelines
[6]https://icml.cc/Conferences/2022/ReviewerTutorial

Dataset	# Purpose	# Review	Avg. length of review (terms of words)	Avg. length of review (terms of sentences)	# Annotated sentences
ICLR 2018	For proposed	1322	345.878	17.511	23150

Table 1: Dataset statistics

Overall (OAL), Bibliography (BIB) and External (EXT).

Paper Aspect Category: The paper aspect category identifies the aspect of the paper that the review-statement addresses. E.g, Appropriateness (APR), Originality or Novelty (NOV), Significance or Impact (IMP), Meaningful Comparison (CMP), Presentation & Formatting (PNF), Recommendation (REC), Empirical & Theoretical Soundness (EMP), Substance (SUB) and Clarity (CLA).

Reviewer - Centric Uncertainty: In peer review, reviewers sometimes make superficial, speculative comments, which are not very helpful, and ultimately affect the outcome (Ghosal et al., 2022b; Özgür and Radev, 2009). For example, some reviewers use vague or hedge words (e.g., maybe, seems, might, etc.) when uncertain about their review. There could be discrepancies between how reviewers comment on themselves and how readers see their preview text. This intuition suggests that a good review will have less reviewer-centric uncertainty (low hedge score). Therefore, we incorporate reviewer-centric uncertainty into our proposed method.

Informativeness Score: Reviews that cover the complete work are more likely helpful to the author (Kühne et al., 2010). It can be an indication of how detailed and significant the judgment was with this intuition. We identify the study corresponding to the paper section and aspects within reviews. The main idea is to arrive at a justifiable informativeness score; if a review is good, it will cover as many sections and important aspects as possible. With this objective, we encoded the annotation label into a numerical score based on the review's coverage across section correspondence, aspect category and reviewer-centric uncertainty of the review by measuring the informativeness score towards the automatic judgment of review quality. We have calculated the informativeness score by considering following three parameters.

3.1.1 i) Section Score ($R_{\mathbf{sec}}$):

A good review should comment on the important sections of the paper, which may help us identify whether the reviewer's comments are semantically related to the submission's main contents. With this intuition, we calculate the section score by given formula.

$$R_{sec} = \frac{\sum \bar{x}_i + \sum \mu_i W_{xi}}{\sum x_i} \tag{1}$$

Where $\Sigma\bar{x}_i$ = no. of unique sections covered by review, μ_{i} = no. of repeating sentences containing i^{th} section, W_{xi} = weight of i^{th} section and $\sum x_i$ = total no. of sections.

3.1.2 ii) Aspect Score ($R_{\mathbf{asp}}$):

As per the rubrics defined (Yuan et al., 2021; Ghosal et al., 2022a) in Peer Review Analyze paper, we expect the review to evaluate the work for indicators like novelty, theoretical and empirical soundness of the research methodology, writing, and clarity of the work, impact of the work in a broader academic context, etc. We call these indicators review-level aspects. We calculate aspect score using the following formula.

$$R_{\text{asp}} = \frac{\sum \bar{x}_i + \sum \mu_i w_{\text{xi}}}{\sum x_i} \tag{2}$$

Where $\Sigma\bar{x}_i$ = no. of unique aspects covered, μ_i = no. of repeating sentences containing i^{th} aspect, w_{xi} = weight of i^{th} aspect $\sum x_i$ = total no. of aspects.

3.1.3 Assigning the Weights $\mathbf{W}_{xi}$:

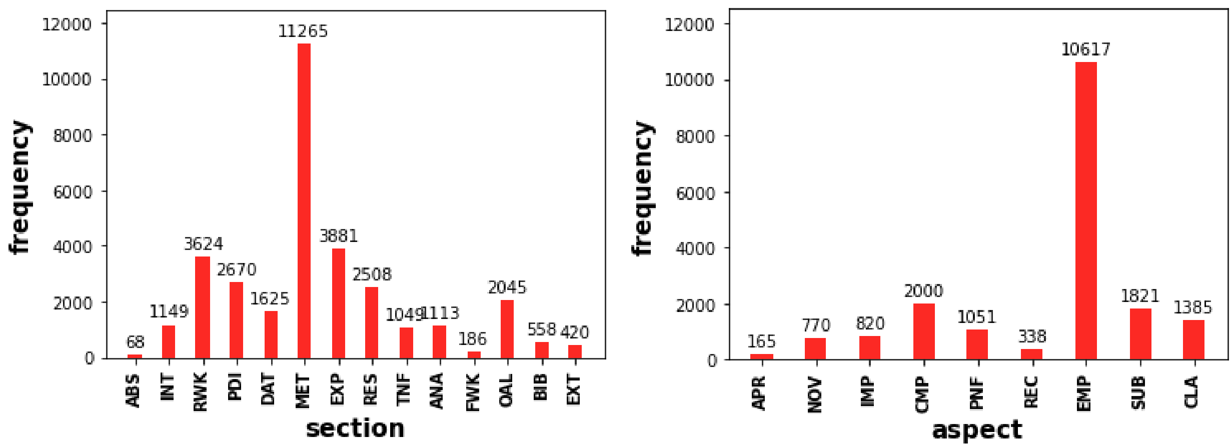
Figure 1 shows the label distribution for each review across the datset for sections and aspects layer. And we assign the weight to respective sections and aspects in our informativeness formula accordingly.

$$W_{xi} = \frac{\text{Freq}_{\text{ xi}}}{100 * \text{Total Freq}} \tag{3}$$

Freq_{xi} = number of sentences talking about a specific section/aspect, Total freq: total number of sentences talking about sections/aspects.

3.1.4 iii) Reviewer-Centric Uncertainty (Hedge Score (H)):

In a review, uncertainty refers to speculation made by the reviewer. The words the reviewer uses to indicate speculating are called hedge words (Lakoff,



(a) Sections distribution (b) Aspect distribution

Figure 1: Sections and aspects distribution across paper section correspondence and paper aspect category in Peer Review Analyze annotated dataset.

1970; Tang et al., 2010; Velldal et al., 2012). Counting uncertain terms in a review is normalized with the number of words in a review to calculate hedge scores. To calculate the hedge score, we use the method proposed by Khandelwal A. et al. (Britto and Khandelwal, 2020; Khandelwal and Sawant, 2019), and we use the XLNet (Yang et al., 2019) version since it outperforms BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019).

$$\text{Hedge Score} = \frac{\Sigma(\text{ hedge words })}{\Sigma(\text{ words })} \tag{4}$$

The score ranges from 0 to 1. If a reviewer is uncertain the hedge score will be higher and vice versa.

Based on the above discussion and using Equations 1, 2, 3 and 4, we derive an informativeness score for a review, which is given below.

$$\textbf{Informativeness score}(R_{\text{info}}) = \frac{R_{\text{sec}}}{e^{H} * e^{1-R_{\text{asp}}}} \tag{5}$$

Where R_{info} = Informativeness score, R_{sec} = Section score, R_{asp} = Aspect score and H = Hedge score.

3.1.5 Intuition about the informativeness score:

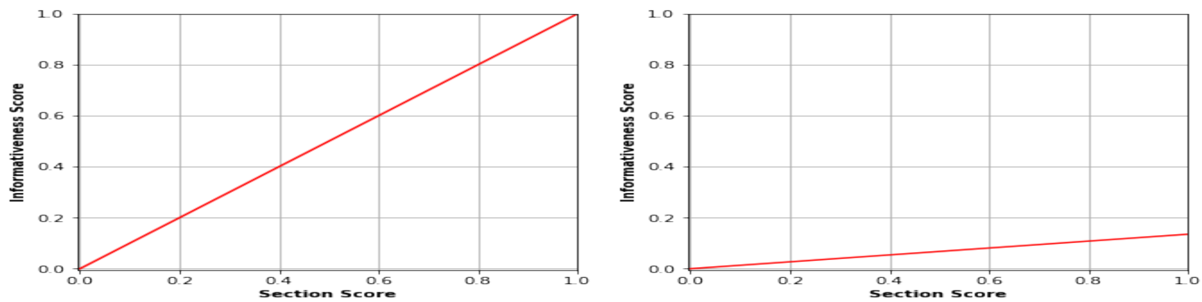
We plot the graph between the informativeness score (R_{info}) and the other three parameters (in the best and worst case). We consider this observation in the informativeness score formula accordingly.

Section Score (R_{sec})**:** From Figure 2, we can see the reason to keep the section score in the numerator.

- Informativeness score is directly proportional to section score $R_{\text{info}} \propto R_{\text{sec}}$ and hence, higher the R_{sec}, higher will be the R_{info} .
- The section score has the highest contribution in determining the informativeness score; as when section score = 0, irrespective of the other two parameters, informativeness score will always be = 0 (see Figure 2.)

Aspect Score ($R_{\textbf{asp}}$)**:** Figure 3 illustrates the relation between informativeness score and aspect score.

- From Figure 3, we can see that higher the aspect score, lower is the $(1 - R_{\text{asp}})$, and hence and value of e^ $(1 - R_{\text{asp}})$ is lower, higher will be the informativeness score. Aspect score has a lower contribution to the informativeness score, as even when aspect score = 0, informativeness score still can be upto 0.3679, depending on the other two parameters (section and hedge score).
- We intend that the informativeness score increases exponentially with increasing aspect score hence, $\mathrm{R}_{\text{info}} \propto \mathrm{e}^{\wedge}\mathrm{R}_{\text{asp}}$. However, to limit the max. R_{info} to 1 at R_{asp} = 1 (Best condition when section score = 1, hedge score = 0) and max. aspect score = 1, we divide the informativeness score by a factor of e.



(a) Best condition (when aspect score = 1 and hedge score = 0) (b) Worst condition (when aspect score = 0 and hedge score = 1)

Figure 2: Informativeness Score Vs. Section Score.

Therefore, $\mathrm{R_{info}} \propto (\mathrm{e}^\wedge\mathrm{R_{asp}})/\mathrm{e}$, which implies that $\mathrm{R_{info}} \propto \mathrm{e}^\wedge(\mathrm{R_{asp}} - 1)$. Hence $\mathrm{R_{info}} \propto 1/\mathrm{e}^\wedge(\mathbf{1} - \mathrm{R_{asp}})$.

Hedge Score (H)**:** Figure 4 illustrates the reason to keep hedge score in the denominator, as a power of e, such that $R_{\mathrm{info}} \propto 1/e^\wedge\mathrm{H}$.

- So, higher the hedge score, higher the $\mathrm{e}^\wedge\mathrm{H}$, and hence lower will be the informativeness score.
- we can see from Figure 4 hedge score has a lower contribution to the informativeness score; as even when hedge score = 1, informativeness score can reach 0.3679, depending on the other two parameters (section and aspect score).
- We intend that the informativeness score decreases exponentially with increasing hedge score, and at $\mathrm{H} = 0$, $\mathrm{R_{info}} = 1$. Hence, $\mathrm{R_{info}} \propto \mathrm{e}^\wedge(-\mathrm{H})$ which implies that $\mathrm{R_{info}} \propto 1/\mathrm{e}^\wedge\mathrm{H}$.

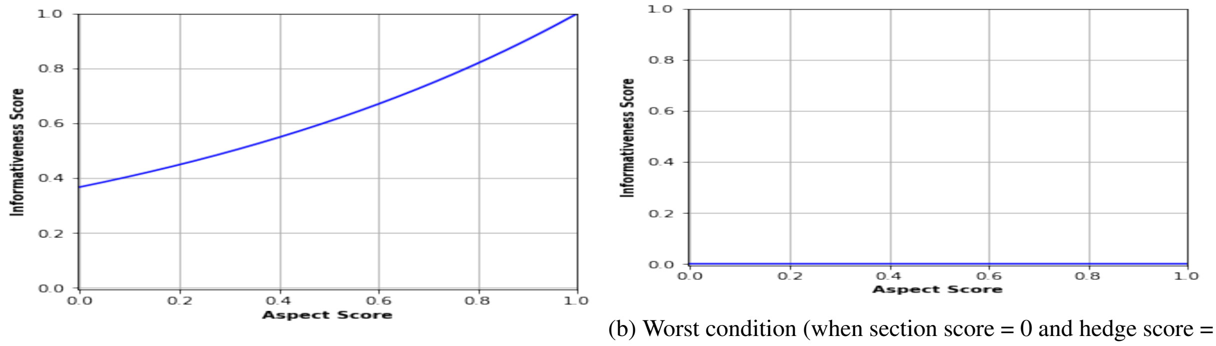
4 Benchmarking Experiments

In addition, we provide baselines for natural language processing (NLP) on the experimental dataset (both annotated and unannotated). Moreover, we train nine methods based on data, including Multiple Linear Regressions (MLR), Robust Regressions (RANSAC), Random Forest Regressions (RF), Long Short-Term Memory (LSTM), Extreme Learning Machines (ELM), Bidirectional Long Short-Term Memory (BiLSTM), Masked and Permuted Pre-training for Language Understanding (MPNet) (Song et al., 2020), Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018), as well as Transformer variants of SciBERT (Beltagy et al., 2019).

4.1 Features for Peer Review Analyze (annotated dataset)

We use a set of features that includes:

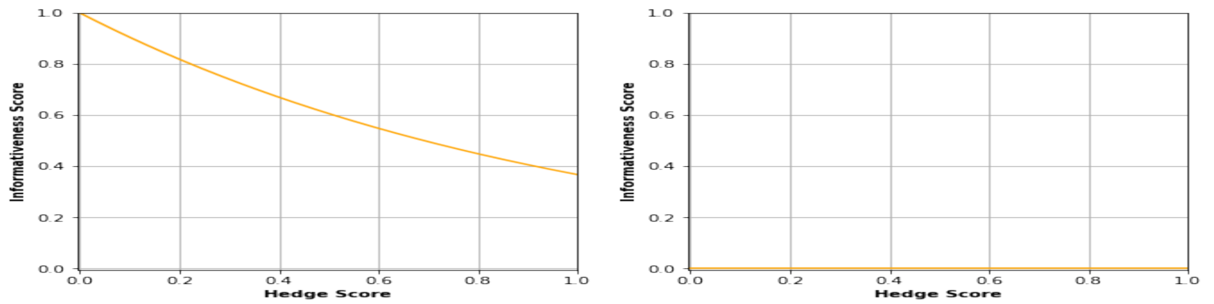
- **Sentence and word count :** We have used the five features sentence count, word count, average sentence length, average word length, and vocab length. The informativeness score is directly proportional to the length of review sentence count and word count, as well as the size of vocabulary vocab length. This gives us a feature matrix of dimension 5.
- **Hedge features:** For review uncertainty, we use the hedge feature hedgescore, which is the average hedge words per sentence, where the hedge words are determined by the method proposed by Khandelwal A. et al. (Britto and Khandelwal, 2020; Khandelwal and Sawant, 2019). This gives us a feature matrix of dimension 1.
- **PoS features:** PoS (Parts of Speech) includes nouns, adjectives, verbs, and adverbs.
- **Sentiment features:** We use VADER (Valence Aware Dictionary for Sentiment Reasoning) (Hutto and Gilbert, 2014) compound sentiment score as the sentiment feature. It ranges from -1 to 1 and gives a feature matrix of dimension 1.
- **Keyword count:** We take the 50 most appearing terms from the papers with top 20% informativeness score as keywords, hence obtaining a feature matrix of 50.
- **Section and aspect coverage:** We use the number of sections covered (out of 14) and the number of aspects covered (out of 9), by



(a) Best condition (when section score = 1 and hedge score = 0)

(b) Worst condition (when section score = 0 and hedge score = 1)

Figure 3: Informativeness Score Vs. Aspect Score



(a) Best condition (when aspect score = 1 and aspect score = 1)

(b) Worst condition (when aspect score = 0 and aspect score = 0)

Figure 4: Informativeness Score Vs. Hedge Score.

the review as features, with feature matrix dimension 2.

- **Section and aspect distribution:** We take the counts of the number of sentences in the review that talks about each section/aspect as features. This gives us a feature of dimension 23.

4.2 Features for unannotated dataset

We use a set of features, which includes sentence and word counts, sentiment features, PoS (Part of Speech), i.e., nouns, adjectives, verbs, and adverbs, hedge features, and keyword counts. Kindly refer to our GitHub repository for the definition and implementation of our full feature set.

Thus, we use feature matrices of dimension 86 for annotated reviews and dimension 61 for unannotated review text (for both, we use Peer Review Analyze dataset) to predict informativeness scores. In addition, word embeddings of their specific dimensions to deep learning models with the Bidirectional Long Short-Term Memory (BiLSTM) pipeline, we use a standard implementation of machine learning models from sci-kit python library, (Pedregosa et al., 2011) keeping the default parameters fixed for a fair comparison across variations in models and embeddings.

Implementation Details: We use Keras on top of TensorFlow-2.4.1 to build the model. Moreover, we train the model with batch size 32, and Adam optimizer with a weight_decay = $\{1e{-}3\}$ to avoid overfitting, and kept each batch balanced while training. We use fixed set $\{1e-1, 1e-2, 1e-3, 3e-3\}$ to tune the learning rate, and find $\{1e-3\}$ works best in our experimental setup. Please see our repository link in the abstract for further information.

4.3 Experimental Setup

In terms of our experimental setup, we use more than one evaluation metrics to avoid any confusion. Because different metrics with the same data can produce different values. It is always better to have a combination of metrics-like MAE (Mean absolute error), Root mean square error (RMSE) and coefficient of determination (R^2) to use together and apply the same metric on a different model to see which one produces the best performance.

5 Evaluation Results & Analysis

We report the evaluation results for annotated and unannotated datasets in Table 2 and Table 3. We kept 80% of the data for training and 20% for eval-

Model Types	MAE	RMSE	(R^2)
MLR	0.0205	0.0305	0.9061
RANSAC	0.0201	0.0295	0.9167
RF	0.0181	0.1924	0.9297
LSTM	0.0178	0.0286	0.9331
ELM	0.0171	0.0267	0.9435
BiLSTM	0.0191	0.0219	0.9619
MPNet	0.0162	0.0184	0.9730
BERT	0.0197	0.0229	0.9583
SciBERT	**0.0152**	**0.0171**	**0.9871**

Table 2: Performance comparision for qualitative analysis on annotated dataset in terms of MAE, RMSE and R-squared (R^2).

Model Types	MAE	RMSE	(R^2)
MLR	0.0596	0.0787	0.3212
RANSAC	0.0666	0.0864	0.3276
RF	0.0682	0.0894	0.3656
LSTM	0.0646	0.0935	0.3051
ELM	0.0636	0.0810	0.3787
BiLSTM	0.0659	0.0878	0.3875
MPNet	0.0657	0.0954	0.2767
BERT	0.0711	0.0931	0.3115
SciBERT	**0.0621**	**0.0735**	**0.4155**

Table 3: Performance comparison for qualitative analysis on unannotated dataset in terms of MAE, RMSE and R-squared (R^2).

uation of the models. We experiment with nine data-driven methods: Multiple Linear Regression (MLR), Robust Regression (RANSAC), Random Forest Regression (RF), Long Short-Term Memory (LSTM), Extreme Learning Machines (ELM), Bidirectional Long Short-Term Memory (BiLSTM), Masked and Permuted Pre-training for Language Understanding (MPNet), Bidirectional Long-Short Term Memory (BiLSTM) on Bidirectional Encoder Representations from Transformers (BERT), and a Bidirectional Long-Short Term Memory (BiLSTM) on Transformer variant of SciBERT, to test the proposed proposition. As shown in Table 2 and Table 3, the deep neural model based on SciBERT representations outperforms both annotated and unannotated datasets.

Qualitative Analysis on Baseline Models: Table 4 shows informativeness score calculate by proposed method and automatically generated informativeness score by nine different techniques on a given Neural Information Processing Systems (NeurIPS) reviews. For qualitative analysis, we take our trained models and predict the score on Neural Information Processing Systems (NeurIPS) sample reviews dataset from the open-access platform OpenReview platform[7]. Table 4 shows some examples of them.

5.1 Case Study:

We analyzed the two ICLR reviews qualitatively to support our proposed method. In the review $https://openreview.net/forum?id=B1EA-M-0Z$. We can see that out of 14 sections, the review has covered 8 unique sections, out of 9 aspects, it covers 4 unique aspects, and this review also has a reviewer-centric uncertainty calculated by hedge score. We can see from Figure 5 (a) the following observations.

- If the review has higher coverage in sections and aspects, the higher will be the section and aspect score. It leads to a higher informativeness score.
- If the reviewer-centric uncertainty (hedge score) is high, then informativeness should be low.

$https://openreview.net/forum?id=ByuP8yZRb$, we can see that out of 14 sections, the review has covered only 6 unique sections, and out of 9 aspects, it covers 3 unique aspects, and this review has high reviewer-centric uncertainty calculated by hedge score. The following observations can be seen in Figure 5 (b).

- This review has low coverage in terms of sections and aspects. Due to this, it has a low informativeness score.
- This review has a high reviewer-centric uncertainty in terms of hedge score, leading to a low informativeness score.

In summary, from this case study shown in Figure 5, we can see the efficiency and suitability of the proposed informativeness method.

6 Conclusion and future work

In this paper, we provide an effective solution to automatically estimate the informativeness score

[7]https://openreview.net/

Review Id	(Informativeness score calculate by proposed method)	Informativeness Score Predicted by Baseline Models								
		MLR	RANSA	RF	LSTM	BiLSTM	ELM	MPNet	BERT	SciBERT
URL: https://proceedings.neurips.cc/paper/2018/file/9246444d94f081e3549803b928260f56-Reviews.html										
NIPS_2018_1006__R1	0.1596	0.1108	0.1176	0.1292	0.1316	0.1381	0.1328	0.1398	0.1347	0.1443
NIPS_2018_1006__R2	0.2713	0.1849	0.1989	0.1998	0.2189	0.2191	0.2212	0.2351	0.2479	0.2569
NIPS_2018_1006__R3	0.5053	0.3992	0.4087	0.4097	0.4162	0.4194	0.4276	0.4459	0.4639	0.4752
URL: https://proceedings.neurips.cc/paper/2018/file/e77dbaf6759253c7c6d0efc5690369c7-Reviews.html										
NIPS_2018_443__R1	0.2822	0.1818	0.1884	0.1931	0.2245	0.2279	0.2311	0.2434	0.2496	0.2765
NIPS_2018_443__R2	0.3249	0.2067	0.2107	0.2256	0.2383	0.2338	0.2430	0.2458	0.2961	0.3006
NIPS_2018_443__R3	0.3236	0.2022	0.2308	0.2355	0.2412	0.2443	0.2536	0.2563	0.3038	0.3151

Table 4: Qualitative analysis results for predicting the Informativeness score by baseline models.

https://openreview.net/forum?id=B1EA-M-0Z

This paper presents a new covariance function for Gaussian processes (GPs) that is equivalent to a Bayesian deep neural network with a Gaussian prior on the weights and an infinite width. [[INT, MET], [NOV]] As a result, exact Bayesian inference with a deep neural network can be solved with the standard GP machinery. [[MET]] Pros: The result highlights an interesting relationship between deep nets and Gaussian processes. [[RES], [EMP]] (Although I am unsure about how much of the kernel design had already appeared outside of the GP literature.) [[EXP]] The paper is clear and very well written. [[OAL], [CLA]] The analysis of the phases in the hyperparameter space is interesting and insightful. [[ANA], [EMP]] On the other hand, one of the great assets of GPs is the powerful way to tune their hyperparameters via maximisation of the marginal likelihood but the authors have left this for future work! [[FWK], [IMP]] Cons: Although the computational complexity of computing the covariance matrix is given, no actual computational times are reported in the article. [[EXP], [EMP]] I suggest using the same axis limits for all subplots in Fig 3. [[TNF]]

Section Score: 0.5717649593
Aspect Score: 0.4456883593
Hedge Score: 0.02312138728
Informativeness Score: 0.3209530879

(a) Informativeness score calculated by proposed method

https://openreview.net/forum?id=ByuP8yZRb

The below review addresses the first revision of the paper. [[EXT]] The revised version does address my concerns. [[OAL]] The fact that the paper does not come with substantial theoretical contributions/justification still stands out. [[PDI, MET], [EMP]] The authors present a variant of the adversarial feature learning (AFL) approach by Edwards Storkey. [RWK]] AFL aims to find a data representation that allows to construct a predictive model for target variable Y, and at the same time prevents to build a predictor for sensitive variable S. [[RWK]] The key idea is to solve a minimax problem where the log-likelihood of a model predicting Y is maximized, and the log-likelihood of an adversarial model predicting S is minimized. [[RWK]] The authors suggest the use of multiple adversarial models, which can be interpreted as using an ensemble model instead of a single model. [[MET]] The way the log-likelihoods of the multiple adversarial models are aggregated does not yield a probability distribution as stated in Eq. 2. [[EXP, MET], [EMP]] While there is no requirement to have a distribution here a simple loss term is sufficient the scale of this term differs compared to calibrated log-likelihoods coming from a single adversary.[[EXP,MET], [EMP]] Hence, lambda in Eq. 3 may need to be chosen differently depending on the adversarial model. Without tuning lambda for each method, the empirical experiments seem unfair. [[EXP, MET], [EMP]] This may also explain why, for example, the baseline method with one adversary effectively fails for Opp-L. [[RWK]] A better comparison would be to plot the performance of the predictor of S against the performance of Y for varying lambdas. The area under this curve allows much better to compare the various methods. [[EXP, MET], [CMP]] There are little theoretical contributions. Basically, instead of a single adversarial model e.g., a single-layer NN or a multi-layer NN the authors propose to train multiple adversarial models on different views of the data. [[MET], [EMP]] An alternative interpretation is to use an ensemble learner where each learner is trained on a different (overlapping) feature set.[[MET]] Though, there is no theoretical justification why ensemble learning is expected to better trade-off model capacity and robustness against an adversary.[[MET], [EMP]] Tuning the architecture of the single multi-layer NN adversary might be as good? [[MET], [EMP]] In short, in the current experiments, the trade-off of the predictive performance and the effectiveness of obtaining anonymized representations effectively differs between the compared methods. This renders the comparison unfair.[[RWK,EXP], [CMP]] Given that there is also no theoretical argument why an ensemble approach is expected to perform better,[[MET], [EMP]] I recommend to reject the paper. [[OAL], [REC]].

Section Score: 0.4317855033
Aspect Score: 0.3378041979
Hedge Score: 0.03217821782
Informativeness Score: 0.2156280483

(b) Informativeness score calculated by proposed method

Figure 5: Qualitative analysis on annotated ICLR Reviews.

of review on the shoulder of uncertainty and review coverage (sections and aspects of the paper). For the proposed method, we used a publicly available Peer Review Analyze dataset of peer review texts, manually annotated at the sentence level (13.22k sentences) across two layers: Paper Section Correspondence and Paper Aspect Category. Next, we transform these categorical annotations to derive an informativeness score of the review based on the review's coverage across section correspondence, aspects of the paper, and reviewer-centric uncertainty associated with the review toward the automatic judgment of review quality. We believe that these interpretations can assist the editors in making better editorial decisions.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Prabhat Kumar Bharti, Tirthankar Ghosal, Mayank Agrawal, and Asif Ekbal. 2022a. How confident was your reviewer? estimating reviewer confidence from peer review texts. In *International Workshop on Document Analysis Systems*, pages 126–139. Springer.

Prabhat Kumar Bharti, Asheesh Kumar, Tirthankar Ghosal, Mayank Agrawal, and Asif Ekbal. 2022b. Can a machine generate a meta-review? how far are we? In *International Conference on Text, Speech, and Dialogue*, pages 275–287. Springer.

Prabhat Kumar Bharti, Shashi Ranjan, Tirthankar Ghosal, Mayank Agrawal, and Asif Ekbal. 2021. Peerassist: Leveraging on paper-review interactions to predict peer review decisions. In *International Conference on Asian Digital Libraries*, pages 421–435. Springer.

Elise S Brezis and Aliaksandr Birukou. 2020. Arbitrariness in the peer review process. *Scientometrics*, 123(1):393–411.

Benita Kathleen Britto and Aditya Khandelwal. 2020. Resolving the scope of speculation and negation using transformer-based architectures. *arXiv preprint arXiv:2001.02885*.

Rachel Bruce, Anthony Chauvin, Ludovic Trinquart, Philippe Ravaud, and Isabelle Boutron. 2016. Impact of interventions to improve the quality of peer review of biomedical journals: a systematic review and meta-analysis. *BMC medicine*, 14(1):1–16.

Michael L Callaham, William G Baxt, Joseph F Waeckerle, and Robert L Wears. 1998. Reliability of editors' subjective quality ratings of peer reviews of manuscripts. *Jama*, 280(3):229–231.

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*.

Rafael D'Andrea and James P O'Dwyer. 2017. Can editors save peer review from peer reviewers? *PloS one*, 12(10):e0186111.

Martin Enserink. 2001. Peer review and quality: A dubious connection?

Yang Gao, Steffen Eger, Ilia Kuznetsov, Iryna Gurevych, and Yusuke Miyao. 2019. Does my rebuttal matter? insights from a major nlp conference. *arXiv preprint arXiv:1903.11367*.

Tirthankar Ghosal. 2019. Exploring the implications of artificial intelligence in various aspects of scholarly peer review. *Bull. IEEE Tech. Comm. Digit. Libr*, 15(1).

Tirthankar Ghosal, Sandeep Kumar, Prabhat Kumar Bharti, and Asif Ekbal. 2022a. Peer review analyze: A novel benchmark resource for computational analysis of peer reviews. *Plos one*, 17(1):e0259238.

Tirthankar Ghosal, Kamal Kaushik Varanasi, and Valia Kordoni. 2022b. Hedgepeer: a dataset for uncertainty detection in peer reviews. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, pages 1–5.

Tirthankar Ghosal, Rajeev Verma, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Deepsentipeer: Harnessing sentiment in review texts to recommend peer review decisions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1120–1130.

Ann T Gregory and A Robert Denniss. 2019. Everything you need to know about peer review—the good, the bad and the ugly. *Heart, Lung and Circulation*, 28(8):1148–1153.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8.

Tom Jefferson, Elizabeth Wager, and Frank Davidoff. 2002. Measuring the quality of editorial peer review. *Jama*, 287(21):2786–2790.

Amy C Justice, Mildred K Cho, Margaret A Winker, Jesse A Berlin, Drummond Rennie, Peer Investigators, PEER Investigators, et al. 1998. Does masking author identity improve peer review quality?: A randomized controlled trial. *Jama*, 280(3):240–242.

Aditya Khandelwal and Suraj Sawant. 2019. Negbert: a transfer learning approach for negation detection and scope resolution. *arXiv preprint arXiv:1911.04211*.

Conny Kühne, Klemens Böhm, and Jing Zhi Yue. 2010. Reviewing the reviewers: A study of author perception on peer reviews in computer science. In *6th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom 2010)*, pages 1–8. IEEE.

Asheesh Kumar, Tirthankar Ghosal, and Asif Ekbal. 2021. A deep neural architecture for decision-aware meta-review generation. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 222–225. IEEE.

Sandeep Kumar, Hardik Arora, Tirthankar Ghosal, and Asif Ekbal. 2022. Deepaspeer: towards an aspect-level sentiment controllable framework for decision prediction from academic peer reviews. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, pages 1–11.

George Lakoff. 1970. Linguistics and natural logic. *Synthese*, 22(1):151–271.

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*.

Arzucan Özgür and Dragomir Radev. 2009. Detecting speculations and their scopes in scientific text. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 1398–1407.

F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, and O Grisel. 2011. Blonde, l m. *Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay E*, pages 2825–2830.

Drummond Rennie. 2016. Let's make peer review scientific. *Nature*, 535(7610):31–33.

Mona M Shattell, Peggy Chinn, Sandra P Thomas, and W Richard Cowling III. 2010. Authors' and editors' perspectives on peer review quality in three scholarly nursing journals. *Journal of nursing scholarship*, 42(1):58–65.

Kyle Siler, Kirby Lee, and Lisa Bero. 2015. Measuring the effectiveness of scientific gatekeeping. *Proceedings of the National Academy of Sciences*, 112(2):360–365.

Amanda Sizo, Adriano Lino, Luis Paulo Reis, and Álvaro Rocha. 2019. An overview of assessing the quality of peer review reports of scientific articles. *International Journal of Information Management*, 46:286–293.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

Buzhou Tang, Xiaolong Wang, Xuan Wang, Bo Yuan, and Shixi Fan. 2010. A cascade method for detecting hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning–Shared Task*, pages 13–17.

Susan Van Rooyen. 2001. The evaluation of peer-review quality. *Learned Publishing*, 14(2):85–91.

Susan Van Rooyen, Nick Black, and Fiona Godlee. 1999. Development of the review quality instrument (rqi) for assessing peer reviews of manuscripts. *Journal of clinical epidemiology*, 52(7):625–629.

Erik Velldal, Lilja Øvrelid, Jonathon Read, and Stephan Oepen. 2012. Speculation and negation: Rules, rankers, and the role of syntax. *Computational linguistics*, 38(2):369–410.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2021. Can we automate scientific reviewing? *arXiv preprint arXiv:2102.00176*.