*INTERSPEECH 2020*
October 25–29, 2020, Shanghai, China

# Uncertainty-Aware Machine Support for Paper Reviewing on the Interspeech 2019 Submission Corpus

*Lukas Stappen*[1*], *Georgios Rizos*[2*], *Madina Hasan*[3], *Thomas Hain*[3], *Björn W. Schuller*[1,2]

[1]EIHW – Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2]GLAM – Group on Language, Audio, & Music, Imperial College London, UK
[3]SpandH – Speech and Hearing Research Group, University of Sheffield, UK

stappen@ieee.org, rizos@ieee.org

## Abstract

The evaluation of scientific submissions through peer review is both the most fundamental component of the publication process, as well as the most frequently criticised and questioned. Academic journals and conferences request reviews from multiple reviewers per submission, which an editor, or area chair aggregates into the final acceptance decision. Reviewers are often in disagreement due to varying levels of domain expertise, confidence, levels of motivation, as well as due to the heavy workload and the different interpretations by the reviewers of the score scale. Herein, we explore the possibility of a computational decision support tool for the editor, based on Natural Language Processing, that offers an additional aggregated recommendation. We provide a comparative study of state-of-the-art text modelling methods on the newly crafted, largest review dataset of its kind based on Interspeech 2019, and we are the first to explore uncertainty-aware methods (soft labels, quantile regression) to address the subjectivity inherent in this problem.
**Index Terms**: natural language understanding, peer review, subjectivity quantification, meta-research

## 1. Introduction

Peer review is one of the cornerstones of the scientific process for presenting, evaluating, and discussing novel findings since the 17th century [1]. Besides the purely written documentation of this process in journals, conferences provide a breeding ground for academic dispute and exchanges with field peers. Acceptance of academic papers for leading conferences is based on the evaluation of the scientific merit thereof, made by the expert reviewers as well as the conference chairs. Peer-review is detailed work, requiring assessment of the article with respect to criteria such as novelty, technical correctness, reproducibility, and thematic alignment with the venue. This is exacerbated by the detrimental effect of numerous reviewer biases to the integrity of the reviews [2]. The individual biases are traditionally addressed by requesting multiple reviewers to evaluate each submission; even so, this occasionally has led to many valuable papers to at least one rejection prior to final acceptance [3, 4].

The impact of such biases on the peer review process can also be surmised from *the NeurIPS experiment* [5], in which 10 % of submissions were put through peer-review twice under different committees, with a 57 % disagreement on the list of accepted papers between committees. Whereas it was observed in [4] that more reviewers per submission lead to a more accurate screening process, this comes at the cost of greater workloads. Instead, we follow the assumption of recent exploratory studies [6–8] that **a Natural Language Processing (NLP) support tool can provide an additional computational recommendation by aggregating the reviews.** In [6], the authors provide *PeerRead*, a dataset of peer-reviews of Artificial Intelligence conferences, and a set of simple NLP baselines by utilising the abstract as well as 1 608 (ICLR '17, ACL '17, CoNLL '16) reviews for predicting submission acceptance. The studies performed in [7, 8] proposed models that predict the submission acceptance based on dual-instance models that pair each review with the corresponding abstract. The former study utilises a Convolutional and Recurrent neural network layer stack that processes classical word embeddings [9], whereas the latter utilise sentence embeddings [10] instead, with additional sentiment metadata extracted by the VADER model [11]. **We are the first, however, that propose a model that examines submission level recommendation by aggregating the review text from all reviewers, in an uncertainty-aware framework that also takes into account the review task subjectivity**, and evaluate it on a corpus of submissions from the world's largest and most comprehensive conference on the science and technology of spoken language processing: the 20th Annual Conference of the International Speech Communication Association – Interspeech 2019[1], promoted as the "*Crossroads of Speech and Language*". Comprising more than 2 000 submissions and around 6 000 reviews, it is the **biggest single-blind, peer review dataset** that has been utilised in such an NLP study.

We explore both submission- and review-level predictions on both a binary classification task of the overall acceptance of a paper, as well as numerical prediction of the individual and summary scores of reviewers. By fusing representations of review texts, we achieve a Macro-F1 score of 57.15% for acceptance prediction using an ALBERT model [12] fine-tuned in an end-to-end manner on our dataset, and a root mean square error of 0.56 ($\pm$ 0.07) for the Interspeech-specific, averaged reviewer overall score using a Convolutional Recurrent Neural Network (C-RNN) trained on FastText word embeddings [13].

## 2. Methods and Experimental Settings

In this section we introduce our methodology, beginning with an explanation of the machine learning tasks and the means by which we propose to address the task subjectivity, followed by the two different model architectures in our experiments: a state-of-the-art ALBERT model fine-tuned on the dataset (cf. Section 2.2), and a more traditional C-RNN that learns from

*Equal Contribution

[1]https://interspeech2019.org/: Due to the high sensitivity of the data, especially with regard to the rejected, and thus still unpublished papers, we are not able to publish the data itself, but the framework code can be found here https://github.com/glam-imperial/Uncertainty-Aware-Machine-Paper-Reviewing.



http://dx.doi.org/10.21437/Interspeech.2020-2862

pre-trained, static word embeddings (cf. Section 2.3). We then describe the review aggregation approaches we adopt for making submission related predictions (cf. Section 2.4). Finally, experimental settings are detailed.

### 2.1. Peer Review Modelling Tasks

In this study, we are concerned with two prediction settings of different granularity: a) *review-level,* in which we make predictions for each individual review based on the review text, and b) *submission-level,* in which we aggregate the reviewers' text information, thus computationally mimicking the job of an editor. We also consider two prediction task types: a) *binary classification* of paper acceptance versus rejection, and b) *regression,* in which we explore the prediction of reviewer scores. Regardless of the prediction setting, the data is expressed as $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$, where $\mathcal{X} = \{X_i\}$ is the set of samples denoted by the index $i$. Similarly, $\mathcal{Y} = \{Y_i\}$, and $Y_i = \{y_{i,s}\}$, where $s$ is the score index (e. g., for a particular submission, we have available the binary acceptance scores, the novelty score, and the technical correctness score, among others). A model $M$ receives the tokenised representation of $X_i$ and learns to produce the predicted numerical values $\hat{y}_{(i,s)}$ for the $s$-th target, which is either a continuous-valued number in the case of regression, or logits in the case of classification.

**Classification (C):** For binary classification, all models optimise the parameters based on the cross-entropy loss. *One of the contributions of this paper is the incorporation of reviewer disagreement information in the training process* by utilising a soft ground truth for each submission that is generated by the different opinions of the reviewers, instead of training the model on the hard binary label. We use the *modified soft label* [14]:

$$q_k = \frac{\alpha + \sum_r^R h_k^r}{2\alpha + \sum_k^{acc.,rej.} \sum_r^R h_k^r}, \tag{1}$$

where $q_k$ is the soft value for the $k$-th class, $h_k^r$ a hard binary value given by the $r$-th reviewer (of a total of $R$ for the submission), and $\alpha$ a regularisation parameter that if set to 0, the standard soft labels are recovered, i. e., the average of the one-hot hard labels. We also utilise the modified version by setting $\alpha$ to $0.75$, following [14]. *Training on soft labels allows the model to place less certainty on submissions towards which the reviewers have exhibited disagreement*. This principle has also been shown to be a successful technique in knowledge distillation [15]. Soft labels are used during training only, to provide information regarding the task subjectivity and the reviewer disagreement to the model; test reporting is performed as usual, on the hard labels, in order to ensure a fair comparison.

**Regression (R):** We further attempt to predict the real valued reviewer scores and use the Mean Squared Error (MSE) as a network training loss. Neural network regression outputs only a point estimate. However, for decision making, it is often beneficial to rely on quantile based confidence intervals to quantify (un)certainty [16–18]. In our context, since a) the reviewers often have disagreeing views on paper merit, and b) in submission-level predictions the final score is only partially dependent on the review, as the final decision is made by the area chair, we believe that it is sensible to be able to model a more complex, distributional output. This can be formalised as:

$$L\left(\xi_i|\alpha_p\right) = \begin{cases} \alpha_p \xi_i & \text{if } \xi_i \geq 0 \\ (\alpha_p - 1)\xi_i & \text{if } \xi_i < 0 \end{cases}, \tag{2}$$

where $\alpha$ is required to be a value between 0 and 1 expressing the percentile $p$ (e. g., $p = 0.9$, the 90th percentile) and $\xi_i = y_i - \hat{y}_{i,p}$ for each individual percentile prediction $\hat{y}_p$. The loss is averaged and updated per batch. When $p = 0.5$, the loss corresponds to the Mean Absolute Error (MAE) except using a median prediction instead of the mean. We implemented **Quantile Regression** (QR) as suggested by [19] by combining multiple-outputs of a predictive model, each corresponding to a different pre-set quantile, to avoid the occurrence of quantile crossing problems.

### 2.2. Transformer Language Models – ALBERT

Recently, a new generation of bidirectional context word embeddings based on Transformer networks [20] has led to significant improvements on several NLP benchmark datasets [21, 22]. NLP-specific Transformer Language Models (TLM), such as BERT [23], XLM [24], RoBERTa [25], and the lightweight, specialised for fine-tuning ALBERT [12] integrate the context of a word by calculating a unique embedding vector for each word based on bidirectional context incorporation. Apart from the training method, the amount of training data and the number of parameters strongly vary between different TLMs. For example, the first bidirectional TLM BERT used the BooksCorpus [26] and the English Wikipedia corpora (combined 16 GBs raw text), resulting in a 110 M parameter network. In comparison, RoBERTa, which is trained on four datasets in total, consists of 160 GBs of source text, and the cross-lingual version XLM utilises 665 M parameters. All have in common that they are very computationally intensive, compared to traditional word embedding methods, for which precalculated versions are common.

We opted to use the ALBERT architecture as a very competitive benchmark model, to perform end-to-end training of a TLM as a standalone network. Compared to the BERT architecture, ALBERT applies two novel parameter reduction techniques: First, the embedding matrix is split into two smaller matrices, and second, layers are divided into groups and used repeatedly. Furthermore, it introduces a new self-supervised loss function that improves handling of multiple sentence input in fine-tuning tasks. These changes reduced the memory footprint, increased training speed and achieved the best results to date in the NLP benchmarks LUE, RACE, and SQuAD, while pre-trained on the same sources (3.3 B words) as BERT.

### 2.3. Word Embeddings + Convolutional Neural Networks

Since TLMs are very resource hungry, we also utilise unsupervised word embedding techniques, such as Word2Vec [9, 27], Global Vectors (GloVe) [28], and FastText [13, 29] that substitute the word tokens with a vector embedded in a latent semantic space, such that similar words imply smaller vector distance. All used conventional word embeddings are pre-trained on various data sets and compressed to 300 dimensions. FastText and GloVe are based on the Common Crawl (1.9 M unique words, 840 B tokens) and Word2Vec on the GoogleNews (3 M unique words, 100 B total) dataset. We additionally experimented with training embeddings exclusively on our data, but the performance was much worse than the pre-trained embeddings.

After the word embedding layer, the transformed text is further processed by a deep neural network. Convolutional Neural Networks (CNNs) have been used successfully for sentence classification [30], and more recently, C-RNNs have been adopted for NLP tasks such as entity recognition [31], as well as binary [32], and multiclass online hate speech classification [33]. Following the architecture used in [32, 33], ours consists of two stacked one-dimensional CNN layers, followed by one Gated Recurrent Unit (GRU) RNN layer. The number of filters and widths of the convolutional layers are 128-128 and 4-4, respectively, and each one is followed by a max pooling layer that

undersamples at a rate and stride of 2-2. The hidden units of the RNN layer are equal to 100. This deep model outputs a hidden state sequence of a length equal to our selected padded max size $T$. Each hidden state is passed through a linear fully connected (FCNN) layer, and we denote the output sequence by $\{h_t\}$. Since the scores we are trying to predict are per sample, we pool the information of the hidden states into a single one using attention-based pooling, inspired by [34]. Specifically:

$$h_{pooled} = \sum_{t=1}^{T} a_t h_t, \tag{3}$$

where $a_i$ are real valued attention weights – one per hidden state in the sequence. In order to calculate them, we use an additional linear FCNN layer with weight matrix $w$, and as such, the attention weights are learnt and calculated via softmax:

$$a_t = \frac{\exp(w^\intercal h_t)}{\sum_{t'}^{T} \exp(w^\intercal h_{t'})}. \tag{4}$$

An additional FCNN processes the pooled hidden state to yield the model output representation.

### 2.4. Review aggregation

In the case of submission-level prediction, if we wish to include the review text information, we are faced with the following problem: *we have multiple reviews, their number ranging from one to five, and we need a way of fusing their information such that the hidden state we propagate in the neural network is of fixed size.* The way we aggregate multiple reviews for ALBERT, is by concatenating them, adding a <NEXT> token between reviews. Even though this induces an ordering among them, we found it is an effective method, that does not require much more overhead during the already computationally heavy end-to-end fine-tuning of the TLM-based ALBERT. In the C-RNN case, we have an alternative that does not introduce a notion of sequence ordering among them. We opted to address this issue using fusion via an attention mechanism, similar to the one described for attention pooling of sequences in the above.

### 2.5. Experimental settings

We have split our dataset into train-validation-test partitions of 50-25-25 percentages. We optimise the model parameters using the train partition, we evaluate after each epoch on the validation partition, and we report test measures using the model that yielded the best validation performance. For the classification task, we report the **Macro-averaged** *F1* score, and use it for monitoring the best validation performance. For further analysis of the skill of the classifier, we also report the **Macro-averaged** *Area Under Receiver Operator Characteristic Curve (AU-ROC)*, *Area Under Precision-Recall (AU-PR)*. For the regression task, we report *RMSE*, and *MAE* and monitor the former.

Following the authors' recommendation, ALBERT is fine-tuned for 5 epochs using a learning rate of *1e-5*, a warmup ratio of *0.06*, and gradient clipping at *1.0*. Due to GPU memory restrictions (32 GBs), we had to use half-precision training and a batch size of 12. Preliminary experiments showed high result stability, since the model is supervised fine-tuned only. For this reason and the high computation requirements, experiments with ALBERT are executed one time. In contrast, our C-RNN has a more complex structure trained from scratch. Hence, we ran our experiments for 40 epochs, with 10 epoch patience, and we report averaged scores across 40 trials. For our C-RNN, hyperparameter search was limited to an evaluation of different learning rates (*1e-3*, *1e-4*, and *1e-5*) and different kernel sizes (*4*, and *25*). Gradient clipping was applied for stability.

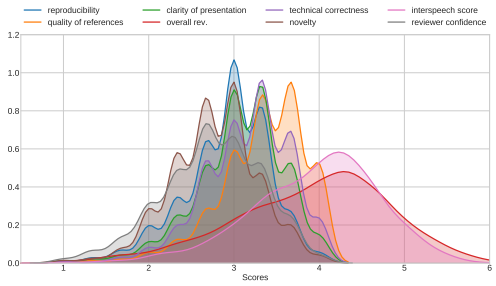


Figure 1: *Density estimation of all scores averaged on submission-level. The two calculated scores, average overall score, and Interspeech score have wider distributions, constituting more difficult tasks.*

## 3. Interspeech 2019 Submission Corpus

The data collected from the submission system of Interspeech 2019 consist of 2 179 preprocessed academic submissions, and 5 842 reviews, with the corresponding acceptance decisions and reviewer score recommendations. Overall, the number of accepted and rejected papers is almost identical. Within the 13 different scientific areas of the conference (tracks), it is still very balanced with less than 5 % deviation. An important criterion for the final acceptance decision, and thus an appropriate regression target, is the aggregate Interspeech score (**iss**):

$$\mathbf{iss} = 1 + 5 * \Big( \frac{\mathbf{conf} + 26}{29} + 0.1 * \frac{\mathbf{nov} - 1}{3} + 0.05 * \frac{\mathbf{tech} - 1}{3} + 0.05 * \frac{\mathbf{pres} - 1}{3} + 0.05 * \frac{\mathbf{rep} - 1}{3} + 0.65 * \frac{\mathbf{reco} - 1}{5} + 0.1 * \frac{\mathbf{qref} - 1}{3} \Big), \tag{5}$$

which is derived from the basic scores of each author, namely a) reviewer confidence (**conf**), b) novelty (**nov**), c) technical correctness (**tech**), d) clarity of presentation (**pres**), e) reproducibility (**rep**), f) overall recommendation (**reco**), and g) quality of references (**qref**) depicted as density distributions in Figure 1.

### 3.1. Data Preprocessing

For the partition sampling, we froze the seed to keep the split constant and balanced over multiple runs, so that at the submission level, $50.8\%$ of the training, $50.3\%$ of the validation, and $48.7\%$ of the test partition contain rejected papers. Small editorial mistakes are corrected, contractions are replaced by full words (you're$\rightarrow$ you are), special characters unified (e. g., – (en-dash) or — (em-dash) $\rightarrow$ - (hyphen)) and transformed to words ( "?" $\rightarrow$ "questionmark") as well as the text is transformed to lower case before the tokenisation. Since the reviews are written in a formal language, no complex preparation was necessary. An average review has 106 words, so that the sequence length was limited to spacious 200 words on review-level and for attention fusion, but increased it by 100 for the review concatenation.

## 4. Results and Discussion

The results for acceptance classification both for submission- and review-level, including the soft label training variations, are summarised in Table 1. Our reviewer score prediction experiment is summarised in Table 2. Finally, we also apply neural network quantile regression in the submission-level setting. The results are summarised in Table 3. We use **bold** font to indicate the performance measure value yielded by the best method for a particular score. In the case of quantile regression, we also denote by *italics* cases where it is better than regular regression.

Table 1: *Performance of Word2Vec (W2V), FastText and Glove (GloVe) word embeddings in our architecture compared to end-to-end fine-tuned ALBERT on review- and submission-level acceptance prediction as hard, soft, and modified soft classification [14]. Classification performance measures are Macro-averaged (%). We report **Mean** (Ø) and standard deviation (±) if multiple trials have been performed.*

| | Measure | W2V Ø | W2V ± | FT Ø | FT ± | GloVe Ø | GloVe ± | Albert Ø |
|---|---|---|---|---|---|---|---|---|
| | | **Review-level** | | | | | | |
| **Hard** | F1 | 58.18 | .05 | **61.62** | .05 | 52.52 | .02 | 56.60 |
| | AU-ROC | 62.98 | .06 | **64.79** | .05 | 54.31 | .02 | 56.60 |
| | AU-PR | 62.04 | .05 | **62.92** | .04 | 54.30 | .01 | 52.23 |
| | | **Submission-level** | | | | | | |
| **Hard** | F1 | 49.23 | .13 | **52.05** | .13 | 45.67 | .11 | 51.37 |
| | AU-ROC | 59.48 | .09 | **61.38** | .10 | 55.70 | .05 | 56.52 |
| | AU-PR | 59.08 | .08 | **61.08** | .10 | 55.31 | .05 | 51.88 |
| **Soft** | F1 | 51.84 | .14 | 55.01 | .13 | 52.83 | .14 | **57.15** |
| | AU-ROC | 62.70 | .09 | **65.22** | .10 | 62.84 | .09 | 57.17 |
| | AU-PR | 62.80 | .09 | **65.86** | .10 | 62.40 | .09 | 52.46 |
| **M.-Soft** | F1 | 46.91 | .14 | **53.86** | .14 | 47.26 | .15 | 32.34 |
| | AU-ROC | 59.80 | .10 | **65.34** | .11 | 59.61 | .09 | 49.43 |
| | AU-PR | 59.82 | .10 | **65.66** | .11 | 59.57 | .09 | 48.07 |

### 4.1. Discussion – classification

On review-level, our framework using FastText embeddings and attention-fusion outperformed all other conventional embeddings, and even the Transformer-based ALBERT on all measures. On submission-level, the results are more varied. *ALBERT yields the best performance in terms of F1 using the regular soft labels, whereas FastText is the best performer in terms of AU-ROC and AU-PR*. The latter is a possible indication that the FastText representations lead to a better calibrated model for this task. The *regular soft labels are consistently contributing to improvement*, especially in the sharp increase observed in the case of ALBERT. The modified soft labels [14] yield a more modest performance improvement in most cases, indicating that the additional degree of softness is not helpful in this case. The GloVe embeddings, which were the representation of choice in the related study performed in [6], were the worst performers in most cases. The TLM-based approach is not the clear winner, as would be expected [23], **perhaps due to the different nature of scientific language, on which these models are not trained, as well as the small, for fine-tuning, size of the dataset.**

### 4.2. Discussion – regression

The scores are predictable to a degree, in accordance to [6], as the RMSE and MAE values are between 0.40 and around 1.00, which is within one reviewer decision level. FastText consistently outperforms other methods in both RMSE and MAE, both for the submission and review level experiments; however, all methods are competitive, and in fact, occasionally the winners, for certain scores. We note that, in general, novelty, technical correctness, and quality of presentation are the easiest scores to predict. Inversely, the overall recommendation, the aggregated Interspeech score, and the reviewer confidence scores are the hardest, which is understandable as they are sampled from the distributions with the widest support, as is depicted in Figure 1.

### 4.3. Discussion – quantile regression

Since the previous experiments have not shown a clear advantage in the usage of the computationally heavier ALBERT method, we opted to perform these experiments on the classical embeddings only. Quantile regression is *more robust*, leading to smaller standard deviations in most cases. This is to be expected, due to the task subjectivity, as a regular regression model would strive to adapt its parameters in order to fit all targets, no matter how unlikely. Quantile regression instead, places less importance on samples it considers to be outliers. Moreover, it is relatively competitive with respect to regular regression, and the better choice in certain cases. Quantile regression seems to be a promising alternative, albeit not a clear improvement.

Table 2: *Prediction of reviewer scores on review-level, as well as the average across reviewers for submission-level prediction. We report **Mean** (Ø) and standard deviation (±) using RMSE (R) and MAE (M) from conventional regression.*

| | Score | W2V Ø R | W2V ± | W2V Ø M | W2V ± | FT Ø R | FT ± | FT Ø M | FT ± | GloVe Ø R | GloVe ± | GloVe Ø M | GloVe ± | Albert Ø R | Albert Ø M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Review-level** | **iss** | 0.88 | .06 | 0.69 | .06 | **0.86** | .03 | 0.69 | .03 | 0.95 | .07 | 0.75 | .07 | 0.78 | **0.61** |
| | **conf** | **0.92** | .01 | 0.72 | .01 | **0.92** | .01 | **0.71** | .01 | **0.92** | .00 | 0.73 | .01 | 0.94 | 0.78 |
| | **nov** | 0.73 | .02 | 0.58 | .02 | **0.71** | .01 | 0.57 | .01 | 0.74 | .02 | 0.57 | .03 | 0.74 | **0.55** |
| | **tech** | **0.69** | .01 | **0.56** | .02 | 0.70 | .01 | 0.58 | .01 | 0.70 | .01 | 0.57 | .01 | 0.73 | 0.60 |
| | **pres** | **0.64** | .02 | **0.50** | .02 | 0.66 | .02 | 0.51 | .02 | 0.65 | .02 | 0.51 | .03 | 0.76 | 0.60 |
| | **rep** | **0.73** | .01 | 0.53 | .01 | **0.73** | .01 | 0.53 | .00 | 0.74 | .01 | 0.53 | .01 | 0.77 | **0.51** |
| | **reco** | 1.09 | .06 | 0.87 | .06 | 1.03 | .02 | 0.82 | .02 | 1.11 | .01 | 0.87 | .01 | **0.97** | **0.74** |
| | **qref** | 0.78 | .01 | **0.64** | .01 | **0.77** | .01 | **0.64** | .01 | 0.80 | .01 | 0.66 | .02 | **0.77** | **0.64** |
| **Submission-level** | **iss** | 0.72 | .02 | 0.57 | .02 | **0.65** | .07 | **0.51** | .06 | 0.70 | .04 | 0.56 | .04 | 0.87 | 0.65 |
| | **conf** | 0.82 | .46 | 0.71 | .46 | **0.59** | .01 | **0.48** | .01 | 0.60 | .01 | **0.48** | .01 | 0.72 | 0.54 |
| | **nov** | 0.49 | .02 | 0.39 | .01 | **0.46** | .03 | **0.37** | .02 | 0.48 | .02 | 0.38 | .01 | 0.51 | 0.40 |
| | **tech** | 0.47 | .01 | 0.38 | .01 | **0.45** | .01 | 0.36 | .01 | 0.47 | .02 | 0.37 | .02 | 0.59 | **0.43** |
| | **pres** | 0.48 | .04 | 0.38 | .03 | **0.44** | .03 | **0.35** | .02 | 0.45 | .03 | 0.36 | .03 | 0.50 | 0.39 |
| | **rep** | 0.51 | .00 | 0.41 | .00 | **0.50** | .00 | 0.41 | .00 | 0.51 | .00 | 0.41 | .01 | 0.50 | **0.40** |
| | **reco** | 0.88 | .03 | 0.70 | .03 | **0.86** | .08 | **0.68** | .06 | 0.87 | .07 | **0.68** | .06 | 0.92 | 0.72 |
| | **qref** | **0.53** | .00 | 0.41 | .00 | **0.53** | .01 | 0.41 | .00 | 0.54 | .00 | 0.41 | .01 | **0.53** | **0.40** |

Table 3: *Prediction of the average scores across reviewers for submission-level prediction using quantile regression. We report **Mean** (Ø) of RMSE (R) and MAE (M) and standard deviation (±).*

| Score | W2V Ø R | W2V ± | W2V Ø M | W2V ± | FT Ø R | FT ± | FT Ø M | FT ± | GloVe Ø R | GloVe ± | GloVe Ø M | GloVe ± |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **iss** | 0.74 | .02 | 0.59 | 0.02 | **0.70** | .03 | **0.56** | .03 | 0.73 | .02 | 0.58 | .02 |
| **conf** | ***0.60*** | .01 | ***0.48*** | 0.01 | **0.60** | .01 | **0.48** | .01 | **0.60** | .01 | **0.48** | .01 |
| **nov** | 0.50 | .00 | 0.40 | 0.00 | **0.48** | .02 | **0.38** | .01 | 0.49 | .01 | 0.39 | .01 |
| **tech** | ***0.46*** | .01 | ***0.37*** | 0.01 | **0.46** | .02 | **0.37** | .02 | 0.48 | .01 | 0.38 | .01 |
| **pres** | 0.50 | .00 | 0.38 | 0.01 | **0.47** | .03 | **0.37** | .02 | 0.50 | .01 | 0.39 | .00 |
| **rep** | ***0.50*** | .00 | ***0.40*** | 0.00 | **0.50** | .01 | ***0.40*** | .01 | 0.51 | .00 | ***0.40*** | .00 |
| **reco** | 0.93 | .00 | 0.73 | 0.01 | **0.92** | .01 | 0.74 | .01 | **0.92** | .02 | **0.72** | .01 |
| **qref** | **0.53** | .00 | ***0.40*** | 0.00 | **0.53** | .01 | 0.41 | .01 | ***0.53*** | .00 | 0.41 | .00 |

## 5. Conclusions and Future Work

We expect that human executive contribution will continue being the major factor in the peer-review process, however, we hope that this study will be an important proof-of-concept towards the design of automated tools that enhance the review process by providing an additional, data-informed recommendation. We took first steps in evaluating the potential of capturing the uncertainty inherent in the task, and found that the usage of soft labels brings clear improvement in submission-level decision classification. We want to emphasise our confidence that uncertainty quantification methods will be crucial in such tasks where the human executive function is traditionally required, and that further research in uncertainty-aware modelling of peer review related data is important. One avenue we consider particularly promising for the future, is the decomposition of the task uncertainty into different factors [35].

## 6. Acknowledgements

This work was partially supported by the UK Economic & Social Research Council through the research Grant No. HJ-253479 (ACLEW). Georgios Rizos was funded by the Imperial College President's Scholarship EPSRC Grant No. 2021037. The authors would also like to thank Imperial College MSc student Korbinian Friedl for his valuable insights regarding the codebase usage.

# 7. References


[1] D. A. Kronick, "Peer review in 18th-century scientific journalism," *Journal of the American Medical Association*, vol. 263, no. 10, pp. 1321–1322, 1990.

[2] D. Chavalarias and J. P. Ioannidis, "Science mapping analysis characterizes 235 biases in biomedical research," *Journal of Clinical Epidemiology*, vol. 63, no. 11, pp. 1205–1215, 2010.

[3] J. S. Gans and G. B. Shepherd, "How are the mighty fallen: Rejected classic articles by leading economists," *Journal of Economic Perspectives*, vol. 8, no. 1, pp. 165–179, 1994.

[4] J. L. Jackson, M. Srinivasan, J. Rea, K. E. Fletcher, and R. L. Kravitz, "The validity of peer review in a general medicine journal," *PloS one*, vol. 6, no. 7, 2011.

[5] N. Lawrence and C. Cortes. (2014) The NIPS experiment. [Online]. Available: http://inverseprobability.com/2014/12/16/the-nips-experiment

[6] D. Kang, W. Ammar, B. Dalvi, M. van Zuylen, S. Kohlmeier, E. Hovy, and R. Schwartz, "A dataset of peer reviews (peerread): Collection, insights and nlp applications," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NNACL, 2018, pp. 1647–1661.

[7] K. Wang and X. Wan, "Sentiment analysis of peer review texts for scholarly papers," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 2018, pp. 175–184.

[8] T. Ghosal, R. Verma, A. Ekbal, and P. Bhattacharyya, "Deepsentipeer: Harnessing sentiment in review texts to recommend peer review decisions," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL, 2019, pp. 1120–1130.

[9] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2013, pp. 3111–3119.

[10] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar *et al.*, "Universal sentence encoder for english," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*. ACL, 2018, pp. 169–174.

[11] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *18th International AAAI Conference on Weblogs and Social Media*. AAAI, 2014.

[12] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," in *International Conference on Learning Representations (ICLR)*, 2019.

[13] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[14] A. Ando, S. Kobashikawa, H. Kamiyama, R. Masumura, Y. Ijima, and Y. Aono, "Soft-target training with ambiguous emotional utterances for dnn-based speech emotion classification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4964–4968.

[15] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[16] G. Ostrovski, W. Dabney, and R. Munos, "Autoregressive quantile networks for generative modeling," in *Proceedings of the 35th International Conference on Machine Learning (ICLM)*, 2018.

[17] W. Dabney, G. Ostrovski, D. Silver, and R. Munos, "Implicit quantile networks for distributional reinforcement learning," in *Proceedings of the 35th International Conference on Machine Learning (ICLM)*, 2018, pp. 1096–1105.

[18] Y. Romano, E. Patterson, and E. J. Candès, "Conformalized quantile regression," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[19] F. Rodrigues and F. C. Pereira, "Beyond expectation: Deep joint mean and quantile regression for spatiotemporal problems," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.

[21] L. Stappen, F. Brunn, and B. Schuller, "Cross-lingual zero-and few-shot hate speech detection utilising frozen transformer language models and axel," *arXiv preprint arXiv:2004.13850*, 2020.

[22] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen *et al.*, "The interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks." ISCA, 2020.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019.

[24] G. Lample and A. Conneau, "Cross-lingual language model pretraining," *arXiv preprint arXiv:1901.07291*, 2019.

[25] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[26] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015, pp. 19–27.

[27] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[28] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[29] L. Stappen, N. Cummins, E.-M. Meßner, H. Baumeister, J. Dineley, and B. Schuller, "Context modelling using hierarchical attention networks for sentiment and self-assessed emotion detection in spoken narratives," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6680–6684.

[30] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, Oct. 2014.

[31] J. P. Chiu and E. Nichols, "Named entity recognition with bidirectional lstm-cnns," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 357–370, 2016.

[32] Z. Zhang, D. Robinson, and J. Tepper, "Detecting hate speech on twitter using a convolution-gru based deep neural network," in *European Semantic Web Conference*. Springer, 2018, pp. 745–760.

[33] G. Rizos, K. Hemker, and B. Schuller, "Augment to prevent: Short-text data augmentation in deep learning for hate-speech classification," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*. ACM, 2019, pp. 991–1000.

[34] Z. Zhang, B. Wu, and B. Schuller, "Attention-augmented end-to-end multi-task learning for emotion prediction from speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6705–6709.

[35] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5574–5584.