15TH IAPR INTERNATIONAL WORKSHOP ON DOCUMENT ANALYSIS SYSTEMS

DAS- 2022 #71

How Confident Was Your Reviewer? Estimating Reviewer Confidence From Peer Review Texts

Presenter : Prabhat Kumar Bharti

Co-authors: Dr. Tirthankar Ghoshal, Dr. Mayank Agrawal, Dr. Asif Ekbal

Indian Institute of Technology Patna - 801103, India

Outline

- Background and Motivation
- Problem Definition
- Our Contributions
- Experimental Dataset
- Proposed Architecture
- Experiments & Results
- Noteworthy findings
- Conclusion and Future work

Background

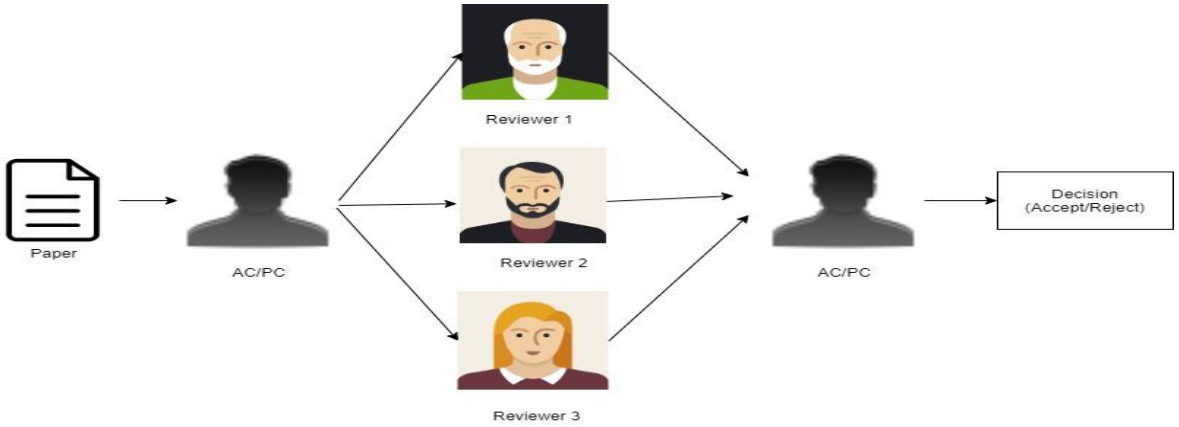
What is peer review system?

- A system to verify and validate a piece of research work before publication. One or more experts review the manuscript before it is published.

- It is followed by the majority of present-day conferences and journals.

How does peer review system work?

- The authors submit their paper for publication.
- A program/area chair will assign reviewers to a research paper.
- Each reviewer reads the article and expresses her opinion on it.



- A program /area chair examines the peer review texts in order to decide whether they should be accepted or rejected.

Motivation

- An ever-increasing volume of research articles being submitted across different venues poses significant managerial challenges for the area/program chairs.
- The quality, randomness, bias, and inconsistency of peer reviews are widely debated within the academic community.
- However, there could be inconsistencies in what reviewers self-annotate themselves versus how the review text appears to the readers.
- Here in this work, we attempt to automatically estimate how confident was the reviewer directly from the review text.

We are curious to see if an AI-based system might ease this burden to some extent by directly predicting a reviewer's conviction based on their review text using Natural Language Processing (NLP) or Machine Learning (ML) techniques?

Problem Definition

There are reviewer set U for paper P and the confidence score matrix $C \in C^{|U| \times |P|}$, where the entry C_{uP} indicates the confidence score of reviewer $u \in U$ towards paper P. For a reviewer u, the review written by u can be represented as $R = \{S_1, \cdots, S_n\}$, for paper P. Where $S_1, \cdots, S_n$ $\rightarrow$ are the sentences in the review texts.

$$f\left(\{R\}_{n=1}^{n}\right) \rightarrow C_{uP}$$

We do not envisage an AI reviewing papers in the near-future, but seek to explore a human-AI collaboration in the decision-making process where the AI would leverage on the human-written reviews to augment human judgment about the quality of a review.

Use Case

A good use case of such an AI would be: assist the editors/program chairs as an additional layer of confidence in the final decision making especially when non conflicting reviews and borderline cases.

Our Contributions

To test this proposition

We experiment with five data-driven methods:

- Linear Regression
- Decision Tree
- Support Vector Regression
- Bidirectional Encoder Representations from Transformers (BERT)
- And a hybrid of Bidirectional Long-Short Term Memory (BiLSTM) and Convolutional Neural Networks (CNN) on Bidirectional Encoder Representations from Transformers (BERT), to predict the confidence score of the reviewer.

- Our experiments show that the deep neural model grounded on BERT representations generates encouraging performance.

Experimental Dataset

- **Source:** https://openreview.net/
- **Ethical Statement:** The reviews from ICLR are publicly available and we crawled using the official OpenReview website.
- For this work, we curate a dataset of 11.5k reviews submitted to ICLR conference and its confidence score for the years 2018, 2019 and 2021 from an open source peer-reviewing portal.

Table 1: Data Statistics and Analysis

Conference Edition	# Reviews	Avg Length of Reviews (in terms of sentences)	Avg Length of Reviews (in terms of words)
2018	2967	22.62	365.12
2019	4764	24.49	394.83
2021	3290	26.23	436.49

Proposed Architecture

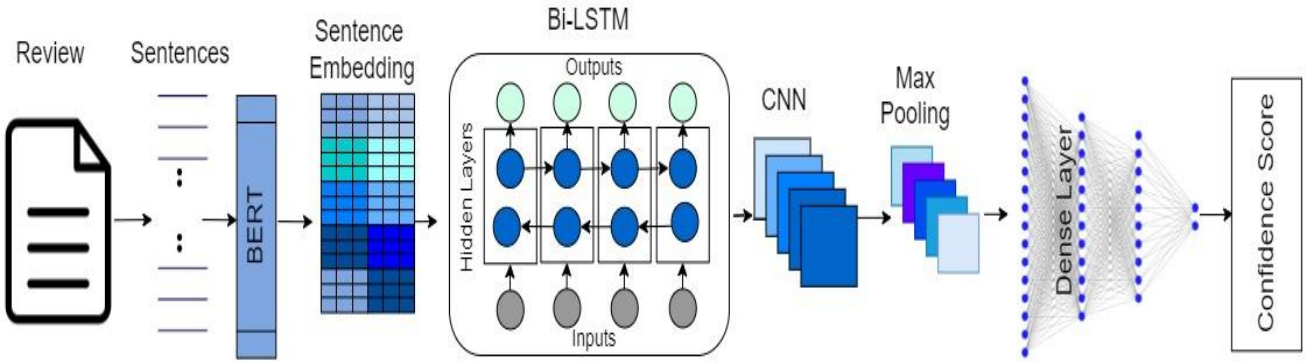


Figure 1: Bidirectional Encoder Representations from Transformers (BERT) using hybrid bidirectional LSTM and CNN architecture for prediction of confidence score

Results

Table: 2 Performance comparisons of linear regression, SVR, decision tree, BERT, and proposed model

Model Types		Conference Edition/ Dataset								
		ICLR 2018			ICLR 2019			ICLR 2021		
		RMSE	MAE	R^2	RMSE	R^2	R^2	RMSE	MAE	R^2
Baselines	Linear Regression	0.944	0.739	-0.358	1.099	0.859	-0.708	1.034	0.828	-0.696
	Decision Tree	1.055	0.767	-0.696	1.116	0.812	-0.759	1.061	0.778	-0.786
	SVR	0.762	0.577	0.115	0.804	0.625	0.085	0.766	0.628	0.045
	BERT	0.591	0.451	0.362	0.617	0.617	0.374	0.605	0.427	0.369
Proposed Model		0.406	0.324	0.689	0.418	0.418	0.654	0.423	0.334	0.6

Ablation Study

- To validates the effectiveness of our proposed framework.

Table: 3 Impact of the proposed model's internal structure

Model Type	Conference Edition/ Dataset								
	ICLR 2018			ICLR 2019			ICLR 2021		
	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2
Proposed Model	0.406	0.324	0.689	0.418	0.323	0.654	0.423	0.334	0.601
Proposed Model w/o BiLSTM	0.536	0.385	0.451	0.514	0.414	0.418	0.495	0.495	0.416
Proposed Model w/o CNN	0.576	0.439	0.395	0.565	0.457	0.382	0.562	0.562	0.386

Cross - Year Experiments

- To test the robustness of our proposed model. We perform cross-year experiments and evaluate the RMSE, MAE, and R^2 scores.

Table 4: Results for cross-year experiments # ICLR means the proposed model is trained on ICLR dataset.

Proposed Model	# ICLR 2018		#ICLR 2019		#ICLR 2021	
	ICLR 2019	ICLR 2021	ICLR 2018	ICLR 2019	ICLR 2018	ICLR 2019
RMSE	0.423	0.443	0.418	0.408	0.403	0.381
MAE	0.334	0.339	0.321	0.309	0.316	0.295
R^2	0.647	0.575	0.626	0.626	0.618	0.638

Noteworthy findings

- This paper proposed a hybrid bidirectional LSTM and CNN architecture grounded on BERT that leverages peer review text as input.
- We compare our studies with Linear Regression (LR), Decision Tree (DT), Support Vector Regression (SVR) and Bidirectional Encoder Representations from Transformers (BERT).
- Our experiments show that the deep neural model grounded on BERT representations generates encouraging performance.
- The proposed model will assist the area or program chair to create an automatic judgment of review quality.
- To evaluate the effectiveness of our proposed framework we perform an ablation study. And we found If either BiLSTM or CNN is removed, we observe the drop in performance across the dataset. These results indicate that both BiLSTM and CNN-based approaches would efficiently guide the framework to make good predictions.
- Additionally, we evaluate our proposed model using data from International Conference on Learning Representations (ICLR) 2018, 2019 and 2021 in a cross-year fashion to verify its efficacy.

Conclusion and Future work

- In this work, we proposed a hybrid BiLSTM and CNN architecture grounded on BERT baseline that leverages review texts to predict the reviewer's confidence score.

- Statistical testing of the proposed model has consistently shown that it outperforms the baselines by a wide margin.

- Whereas we do not envisage an AI to take up the role of a reviewer, but our work could be a step towards human-AI collaboration in peer reviews.

- In the future, we intend to explore how we can broaden the scope of our work by modeling the linguistic properties of the review content as they frame uncertainty and conviction.

Thank You Very Much!!
Open up for Q & A (on-topic questions, please)