What Makes a Scientific Paper be Accepted for Publication?

Panagiotis Fytas[1] **Georgios Rizos**[1] **Lucia Specia**[1, 2]
[1]Department of Computing, Imperial College London, UK
[2]Department of Computer Science, University of Sheffield
{pf2418, georgios.rizos12, l.specia}@imperial.ac.uk

Abstract

Despite peer-reviewing being an essential component of academia since the 1600s, it has repeatedly received criticisms for lack of transparency and consistency. We posit that recent work in machine learning and explainable AI provide tools that enable insights into the decisions from a given peer-review process. We start by simulating the peer-review process using an ML classifier and extracting global explanations in the form of linguistic features that affect the acceptance of a scientific paper for publication on an open peer-review dataset. Second, since such global explanations do not justify causal interpretations, we propose a methodology for detecting confounding effects in natural language and generating explanations, disentangled from textual confounders, in the form of lexicons. Our proposed linguistic explanation methodology indicates the following on a case dataset of ICLR submissions: a) the organising committee follows, for the most part, the recommendations of reviewers, and b) the paper's main characteristics that led to reviewers recommending acceptance for publication are originality, clarity and substance.

1 Introduction

The peer review process has been instrumental in academia for determining which papers meet the quality standards for publication in scientific journals and conferences. However, it has received several criticisms, including inconsistencies among review texts, review scores, and the final acceptance decision (Kravitz et al., 2010), arbitrariness between different reviewer groups (Langford and Guzdial, 2015) as well as reviewer bias in "single-blind" peer reviews (Tomkins et al., 2017).

Explainable AI (XAI) techniques have been shown to provide global understanding of the data behind a model.[1] Assume we build a classifier that predicts whether a paper is accepted for publication based on a peer review. Interpreting the decision of such a classifier can help comprehend what the reviewers value the most on a research paper. XAI allows us to elicit knowledge from the model about the data (Molnar, 2019). However, in order to determine what aspects of a paper lead to its acceptance, we need to interpret the peer review classifier explanations *causally*. Generally, XAI methods do not provide such guarantees since most Machine Learning models simply detect correlations (Molnar, 2019). Nevertheless, in recent years there has been a trend towards enabling Machine Learning to adjust for causal inference (Schölkopf, 2019).

Simply interpreting a non-causal classifier does not suffice when attempting to gain insight into human decision-making in the peer review process. For instance, words such as "gan", and "efficiency" appear to be important in such a classifier (Section 4). We argue that correlation between such words and paper acceptance is confounded on the subject of the paper: different subjects have different probabilities of acceptance, as well as different degrees to which algorithmic "efficiency" is an important factor. Since paper subject ground truth is not necessarily available, we can treat the abstract of a paper as a proxy for its subject, similar to Veitch et al. (2019).

Previous approaches that aim to reconcile NLP with causal inference are limited to either binary treatment variables (Veitch et al., 2019; Roberts et al., 2020; Saha et al., 2019) or nominal confounders (Pryzant et al., 2017, 2018a,b). This paper takes a first step towards using learnt natural language embeddings both as the treatment and the confounder. We achieve this by generating deconfounded lexicons (Pryzant et al., 2018a) through

[1]Among the several definitions for explainability and interpretability (Lipton, 2016; Rudin, 2018; Murdoch et al., 2019), we follow the definition of explainability as a *plausible* justification for the prediction of a model (Rudin, 2018), and use the terms interpretability and explainability interchangeably.

Proceedings of CI+NLP: First Workshop on Causal Inference and NLP, pages 44–60

interpreting causal models that predict the acceptance of a scientific paper, based on their peer reviews and confounded on their abstract. Our **main contributions** are:

- We provide a methodology for developing text classifiers that are able to adjust for confounding effects expressed in natural language. Our evaluation is quantitative, reporting the Informativeness Coefficient measure (Pryzant et al., 2018a,b), and we also showcase the highest scoring words from the lexicons.
- We extend the classifier of Pryzant et al. (2018b) to use a black-box, instead of an interpretable classifier, which can be explained through model-agnostic tools, such as LIME (Ribeiro et al., 2016).
- We utilise our methodology to extract insights about the peer-review process of ICLR 2017, given that certain assumptions hold. Those insights validate our perceptions of how the peer-review process works, indicating that the method provides meaningful reasoning.
- We develop novel models for the task of peer review classification, which achieve a 15.79% absolute increase in accuracy over previous state-of-the-art (Ghosal et al., 2019).

In experiments with a dataset from ICLR 2017, we found that the organising committee mostly followed the recommendations of the reviewers, and that the reviewers focused on aspects such as originality, clarity, impact, and soundness of a paper, when suggesting whether or not the paper should be accepted for publication.

The paper is organised as follows: Section 2 presents related work regarding computational analyses of the peer-review process and explainability in NLP. Section 3 gives an overview of PeerRead, the peer review dataset we used in this paper. Section 4 presents our initial exploration on explanation of peer-reviewing, while Section 5 describes our methodology to account for causality by generating deconfounded lexicons.

2 Related Work

Peer Review Analysis. The PeerRead dataset (Kang et al., 2018) is the only openly available peer review dataset. It contains submission data for arXiv and the NeurIPS, ICLR, ACL, and CoNLL conferences, however, review data for both accepted and rejected submissions exist only for ICLR. The authors devised a series of NLP prediction baselines on the tasks of paper acceptance based on engineered features, as well as reviewer score prediction based on modelling abstract and review texts. Improvements on these baselines have been proposed by using better text representations as well as joint abstract/review modelling for predicting paper acceptance (Ghosal et al., 2019; Wang and Wan, 2018). Stappen et al. (2020) revisited the latter task on a dataset containing all the Interspeech 2019 conference abstracts and reviews (which unfortunately is not available) by utilising a review text fusion mechanism. The aforementioned studies model correlations between textual indices and the desired targets, without attempts towards explainability or identification of causal relationships. Hua et al. (2019), utilise argumentation mining on review texts from the ICLR and UAI conferences. Whereas their methodology focuses on proposition type statistics and transitions, it is a distinct approach to understanding peer-reviewing compared to ours, which is based on XAI.

Interpretability in NLP. Interpreting aspects of the input as conducive to the model prediction aims to: a) increase model trustworthiness, and b) allow for greater understanding of the data (Molnar, 2019). One way to do this is via **local explanations**, which explain the result of a single prediction, such as the view of high, learnt, class-agnostic attention weights (Bahdanau et al., 2015) applied to Recurrent Neural Network (RNN) hidden states, as proxies of word importance. However, the use of attention weights as explanations has been controversial, as Jain and Wallace (2019) have pointed out that for the same prediction, there can be counterfactual attentional explanations. Inversely, Wiegreffe and Pinter (2019) offer the justification that there may indeed exist multiple plausible explanations, out of which the attention mechanism captures one. An additional limitation of the attention mechanism is the inability to provide class-specific insights, due to the use of the softmax activation function. Alternatively, the predictions of black-box classifiers can be interpreted through model-agnostic frameworks, such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017).

We are interested in **global explanations**, which explain either a set of predictions or provide insights about the general behaviour of the model.

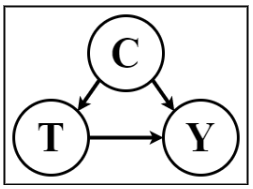

Figure 1: Graphical model depicting the confounding effect: C – Confounder, T – Treatment, Y – Outcome.

Current research is limited either to feature-based classifiers (Altmann et al., 2010), image classifiers (Kim et al., 2017) or combining local explanations to generate global explanations (Lundberg et al., 2019; Ibrahim et al., 2019).

Causal Inference in NLP. Let us assume two observed random variables: the treatment T and the outcome Y. Correlation between those two variables does not necessarily entail causation. Instead, this correlation could be the product of a confounding effect: a third random variable C causes both of the observed random variables (Peters, 2017). A depiction of the confounding effect can be observed in Figure 1. Recently, researchers have provided frameworks for treating text as a confounder, but their approaches are limited to binary treatment variables (Veitch et al., 2019; Roberts et al., 2020; Saha et al., 2019). Keith et al. (2020) have compiled a review of research that utilises text for adjusting for confounding. Fong and Grimmer (2016) explore extracting treatments from text corpora and estimating their causal effect on human decisions. However, they require human annotators for the constructions of training and test sets. Pryzant et al. (2017, 2018a,b) examine the problem of identifying a deconfounded lexicon (a set of linguistic features such as words or n-grams) from the text, which acts as a treatment variable. A deconfounded lexicon is "predictive of a target variable" but "uncorrelated to a set of confounding variables" (Pryzant et al., 2018b). However, their work is limited to using nominal confounders. In this paper, we explore the use of natural language both as the treatment variable and the confounder.

3 Dataset

The PeerRead dataset (Kang et al., 2018) consists of 14.7k papers and 10.7k peer reviews, including the meta-reviews by the editor. The papers are from different venues, having been collected with different methods, which leads to a non-uniform dataset. For instance, arXiv papers do not contain any reviews, and consequently, they are out of scope for our study. Furthermore, the NeurIPS section contains 2k reviews for only accepted papers. Therefore, using them will lead to a significant imbalance in our data. For those reasons, we decided to examine the 1.3k ICLR 2017 reviews of PeerRead. The ICLR 2017 section is divided into training, validation and test sets with an 80%-10%-10% split. In the Appendix, we can observe the proportion of the accepted to rejected paper in the various partitions of ICLR 2017.

PeerRead suffers from various data quality issues, which we have resolved. Firstly, we have removed several empty and duplicate reviews. More importantly, for ICLR 2017 the meta-reviews which contain the review and final decision of the conference chair are not marked as *meta-reviews* (the meta-review boolean field is marked as false) as they should have been (Kang et al., 2018). Instead, they are all marked as normal reviews with the title "ICLR Committee Final Decision". We explicitly treat them as meta-reviews, a differentiation that we believe is crucial in an XAI study like ours. Notably, subsequent studies that utilise PeerRead, like DeepSentiPeer (Ghosal et al., 2019), do not mention whether they have addressed this issue. This hinders a direct comparison with results from DeepSentiPeer as, expectedly, according to our experiments, the use of the meta-reviews significantly increases the performance of classifiers.

4 Data Exploration with Global XAI

We make an initial exploration into global explanations in the form of lexicons by mapping each word to an aggregate of scores corresponding to local explanations of PeerRead test sample predictions, as suggested both by the creators of SHAP (Lundberg et al., 2019) and by Pryzant et al. (2018b). We follow a train-test split setup in our modelling experiments, and only extract explanations from the test set, following Molnar et al. (2020).

We experimented with three different classification tasks: a) the *Final Decision Peer Review Classifier* fuses the multiple peer reviews to predict the final acceptance decision for a paper, b) the *Meta-Review Classifier* uses the singular meta-review to predict the final acceptance decision for a paper, and c) the *Individual Peer Review Classifier* predicts the recommendation of the reviewer for a single review, where we consider as accepted the papers with scores above 5, given a range of 1-10.

4.1 Interpreting Two Classifier Architectures

After preliminary experiments with Transformer Language Models (TLMs), we have opted to use the $\texttt{SciBERT}_{\texttt{SciVocab uncased}}$ model (Beltagy et al., 2019) trained on 1.7M papers from Semantic Scholar (Ammar et al., 2018) that contain a total of 3.17B tokens for text representation.

We experimented with two different text modelling approaches, each allowing for a different local explanation generation technique. In our **first approach**, we used the [CLS] token of our SciBERT model for *review text representation*. We truncate the end of reviews longer than 512 words. The effect of this truncation is not extremely adverse since only 10.5% of the reviews exceed the maximum length. More details about the length of peer reviews are included in the Appendix. We are not fine-tuning our model due to the risk of overfitting on our small dataset of $1.3k$ reviews. For the Meta Review and the Individual Review Classifiers, an additional feed forward neural layer is required, followed by a sigmoid activation function to predict the decisions. In the case of the Final Decision Peer Review Classifier, multiple reviews exist, leading to an equal number of SciBERT representations. In order to avoid introducing an arbitrary ordering among the reviews, we fused them into a single representation using an attention mechanism, following Stappen et al. (2020). We used two more attention layers to produce variant fused embeddings, and concatenated them into a single vector of fixed dimensionality in order to simulate multi-headed attention (Vaswani et al., 2017).

For the **second approach**, we only explore the Meta-Review and Individual Peer Review Classifiers. We treat each review as a sequence of word representations given by SciBERT, which we further process using an RNN layer with a Gated Recurrent Unit cell (GRU), followed by an attention mechanism, which is used to provide interpretability to our model. The output of the RNN layer is a vector produced through attention pooling. This vector is input to a feed-forward layer to produce the classification decision.

For the SciBERT [CLS] approach, we treat our model as a black-box and use **LIME** (Ribeiro et al., 2016) to produce explanations for each test set sample. More important words will have larger absolute scores. For the GRU-based approach, the attention mechanism is used to generate local explanations. More important words will have larger weights. Global explanation scores are then computed as the average of either the LIME score or the attention weights. The implementation details of our model are discussed in the Appendix.

4.2 Results

The validity of the explanations in drawing conclusions about peer reviewing is tied to the model predictive performance (Molnar et al., 2020). Table 1 summarises the performance on the PeerRead held-out test set for the task of predicting the final acceptance recommendation. Although not directly comparable, as explained in Section 3, we also report the baseline performances of PeerRead (Kang et al., 2018) and DeepSentiPeer (Ghosal et al., 2019), as well as that of the Majority Baseline, according to which the prediction is equal to the multiple reviewer recommendation majority vote. In the case of Individual Peer Review classification, this aggregate prediction is the same for all review samples pertaining to the same paper. We achieve superior performance over the baselines, both when using the meta-review and when fusing all the peer reviews for a specific paper. We note that the meta-review based model is the highest performer as the text is expected to correspond very well to the final recommendation. However, even our regular review fusion model outperforms DeepSentiPeer by a 15.79% absolute increase in accuracy. We hypothesise that this happens due to DeepSentiPeer treating each review (even along with the paper) as a different sample; we use fusion. Another reason for this improvement is presumably the use of the SciBERT model which was fine-tuned on biomedical and computer science papers.

Table 2 summarises the results for the Individual Peer Review Classification task. From the above experiments, we see that the SciBERT [CLS] approach is superior to sequential GRU.

In Table 3, we present the top 50 important words in the peer reviews for the Individual Peer Review Classification. In the Appendix, we include explanations for the other models. For both meta-reviews and peer reviews, we observe various technical terms ("lda", "gans", "variational", "generative", "adversarial", "convex", etc.). While this indicates a positive correlation between some subjects (such as GANs) and the acceptance to ICLR 2017, it is hard to argue about a causal relationship. For instance, a possible non-causal relationship could be the following confounding effect: top re-

Model Type	Model	Accuracy	Macro-F1	Weighted-F1
Baselines	Majority Baseline	59.52	0.3731	0.4442
	PeerRead (only paper)	65.30	N/A	N/A
	DeepSentiPeer (paper & review)	71.05	N/A	N/A
Meta-Review	SciBERT	**89.47**	**0.8899**	**0.8947**
	GRU	86.84	0.8661	0.8698
Final Decision Peer Review	SciBERT	86.84	0.8606	0.8676

Table 1: Performance of the classifiers that predict the final decision prediction (i.e., whether a paper is accepted for publication on ICLR 2017) for a paper on the test set of PeerRead. We see that using Meta-reviews in training yields an advantage, as does peer review fusion.

Model Type	Model	Accuracy	Macro-F1	Weighted-F1
Baselines	Majority Baseline	50.76	0.3367	0.3418
	Logistic Regression	58.33	0.5725	0.5756
Individual Peer Review	SciBERT	**80.30**	**0.8026**	**0.8028**
	GRU	70.45	0.7007	0.7012

Table 2: Performance on the Individual Peer Review Prediction (i.e., whether a reviewer suggests that a paper is accepted for publication), using on the test set of PeerRead.

searchers perform novel research, such as GANs, and top researchers have a higher probability of having their work published.

Furthermore, we observe plenty of terms that seem to directly criticise the quality of the work accomplished in a paper: "efficiently", "confusing", "unconvincing", "superficial", "comprehensive", "systematically", "carefully", "untrue" and more. Again, it is easy to make the mistake of assuming a causal relationship between the "efficiency" of an algorithm, suggested in a paper, and the acceptance for publication. For instance, depending on the subject of a paper, the efficiency may or may not increase the chances of a publication. To elaborate, for a paper about real-time language interpretation, the efficiency of the algorithm may directly influence the acceptance decision. On the contrary, the efficiency of training a GAN model may be inconsequential, as far as the ICLR reviewers are concerned.

5 Producing Causal Explanations

"Unjustified Causal Interpretation" is one of the pitfalls of global explanations (Molnar et al., 2020). Simply put, global explanations detect correlation and correlation does not entail causation. In order to learn what leads to a scientific paper being accepted for publication, we need causality. For instance, words such as "efficiency" and "novelty" appear to be important. However, there may not exist a causal relationship between them and the acceptance of a paper. Specifically, more theoretical subjects may demand "novelty", and the chances of a theoretical subject being accepted may be greater. Therefore, the correlation of the novelty and the acceptance of a paper may be an artefact of this confounding effect on the subject of the paper.

Adjusting for every possible confounding and mediating effect is not possible for most machine learning models, partially due to lack of ground truth data. Causality in machine learning is limited to taking several strong assumptions and providing a causal interpretation, for as long as those assumptions hold (Molnar et al., 2020). When text is involved, even more assumptions must be made, due to its high dimensionality (Veitch et al., 2019). We assume that: a) the subject of a paper is sufficient to identify the causal effect, b) the abstract can act as a proxy to the subject, and c) the embedding method extracts information relevant to the subject. In Section 5.4, we discuss the validity and limitations of those assumptions. Even when such assumptions are violated, our methodology is meaningful since it allows disentangling explanations of text classifiers from the natural language embedding of the text we want to control.

5.1 Background

In Figure 1, we depict the confounding effect. In this context, Y is the target variable (e.g. the acceptance of a paper for publication), T is the text (e.g. a peer review), and C are the confounding variables (e.g. the subject of a paper).

The task is to extract a deconfounded lexicon L such that $L(T)$ (the set of words of L that exist in the text T) is correlated to the target variable Y but not to the confounders C (Pryzant et al., 2018a). For our purposes, this means that the extracted

lexicon can offer explanations about the acceptance of a paper for publication regardless of the subject of a paper. Pryzant et al. (2018a) have introduced the concept of *Informativeness Coefficient* $I(L)$:

$$I(L) = \mathbb{E}[\text{Var}[\mathbb{E}[Y|L(T), C]]|C]$$

where $\text{Var}[\mathbb{E}[Y|L(T), C]]$ is the amount of information in the target variable Y that can be explained both by $L(T)$ and C. In order to extract a deconfounded lexicon L, we must maximise the Informativeness Coefficient $I(L)$. The use of $I(L)$ as an evaluation of the quality of causal explanations is motivated by the ability of $I(L)$ to measure the causal effect of the text T on Y, under certain assumptions (Pryzant et al., 2018b).

To further understand this concept, let us consider a lexicon L of fixed length. $L(T)$ contains either words descriptive of $T \setminus C$ or words descriptive of $T \cap C$. Words related to the confounders (from $T \cap C$), in theory, do not increase the Informativeness Coefficient since the confounder C is treated as a fixed variable that directly affects Y. For instance, the word "gan" in a peer review may not be as important when the subject of the paper (the confounder) is already taken into consideration. On the contrary, words like "enjoyable", which are potentially unrelated to the subject of the paper, can contain more information. Therefore, the fewer the words in the lexicon that are related to the confounders C, the more deconfounded a lexicon will be, and $I(L)$ will take larger values.

An ANOVA decomposition is used to write the Informativeness Coefficient of a lexicon $I(L)$ (Pryzant et al., 2018a):

$$\begin{aligned} I(L) = &\mathbb{E}[(Y - \mathbb{E}[Y|C])^2] \\ &- \mathbb{E}[(Y - \mathbb{E}[Y|L(T), C])^2] \end{aligned} \quad (1)$$

In practice, $I(L)$ can be estimated by fitting a logistic regression on the confounders C to predict the outcome Y and a logistic regression on both C and $L(T)$ (a Bag-of-Word created from the lexicon and the text T). Then, for binary classification, the Binary Cross Entropy error is used to measure $I(L)$:

$$I(L) \simeq BCE_C - BCE_{L(T),C}$$

Deep Residualisation (DR) (Pryzant et al., 2018a) is a method for extracting a deconfounded lexicon. A visualisation of the DR model can be seen in Figure 2. The main idea is to utilise a neural model to predict the target variable $\hat{Y}'$ using only **nominal** confounders C, and an **interpretable** model that encodes the text T into an embedding vector e. Then, the intermediate prediction $\hat{Y}'$ is concatenated with the embedding vector e and forwarded through a feed-forward network to produce the final prediction $\hat{Y}$. In order to train this model, two loss functions are used. Firstly, the errors from the intermediate prediction $\hat{Y}'$ (i.e. BCE_C) are propagated through the Intermediate Prediction Neural Network. Secondly, the errors from the final prediction $\hat{Y}$ (i.e. $BCE_{T,C}$) are propagated through the complete neural network (Pryzant et al., 2018a).

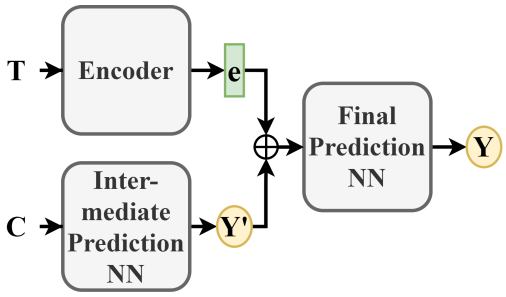

Figure 2: Deep Residualisation Architecture (Pryzant et al., 2018a). The confounders C are used for intermediate predictions Y'. The treatment T (i.e. the review of a paper) is encoded to a vector e, which is concatenated with Y' to produce the final prediction Y.

The DR architecture and the choice of loss functions are directly influenced by the Informativeness Coefficient. Specifically, Equation 1 provides additional intuition behind the Informativeness Coefficient: $I(L)$ "measures the ability of a lexicon $L(T)$ to enhance the predictions of the outcome Y made from the confounders C" (Pryzant et al., 2018b). The DR architecture utilises C to produce intermediate Y' predictions, while the text T is used to improve upon the intermediate predictions (Pryzant et al., 2018b). Intuitively, if the encoder learns to focus on aspects of the text, which are unrelated to the confounders but predictive of the outcome, the improvement of the intermediate predictions, and consequently $I(L)$, will be maximised.

5.2 Submission Subject as a Confounder

Although there may exist multiple confounders to the outcome (e.g., reviewer guidelines, or the paper content itself), we focus specifically on the subject of the paper as a confounder. Such annotation, however, is not part of the available data. Similar to the study performed in Veitch et al. (2019), we assume that the paper abstract is predictive of its subjects,

and opt to utilise it for adjusting for confounding.

A limitation of the work of (Pryzant et al., 2018b) is that they only consider nominal confounders, which they forward through a Multi-layer Perceptron (MLP). We have extended the DR algorithm to extract deconfounded explanations from the reviews while using learnt natural language embeddings as the confounder. Specifically, we are forwarding our confounder (the abstract text) through a SciBERT Language Model to produce an embedding and then, forward that embedding through an MLP, with a sigmoid activation function at the end, to predict the intermediate classification outcome $\hat{Y}'$. For the task of meta-review classification, $\hat{Y}'$ aims to predict the final acceptance outcome, and for the task of individual peer review classification, $\hat{Y}'$ aims to predict the recommendation of the reviewer. Similarly with Veitch et al. (2019), we assume that the embedding method is able to extract information relevant to the subject of the paper.

Firstly, we have extended the **DR+ATTN** and **DR+BoW** models (Pryzant et al., 2018b) to allow for treating of natural language as the confounder, using the aforementioned methodology. The DR+ATTN model utilises an RNN-GRU layer with an attention mechanism to encode the text into the vector e. Therefore, interpretability is achieved through the attention mechanism. The DR+BoW model, forwards a Bag-of-Words (BoW) representation of the text through a single linear neural layer to produce an one-dimensional vector e. Each feature of our model is an n-gram that may appear on the text of the review. Therefore, we can globally interpret our model by using the weight of an n-gram in the single linear layer as the importance of that word (Pryzant et al., 2018b).

Secondly, instead of an interpretable model, we utilise a black-box that encodes the text into the vector e and then, use LIME to explain the predictions of our classifier. Therefore, we have developed the **DR+LIME** variant which can be used to extract a deconfounded lexicon from BERT-based models.

5.3 Quantitative Evaluation

In Table 4, we present metrics for benchmarking the performance of our various lexicons. In order to add more clarity, apart from the Informativeness Coefficient $I(L)$, we include the performance (F1-score) of the different logistic regression models. Furthermore, we report the Informativeness Coefficient of the lexicons generated by the Logistic Regression (LR) and the Logistic Regression with Confound features (LRC) (McNamee, 2005) baselines.

The Informativeness Coefficient is useful for measuring the degree to which the lexicons are deconfounded. However, the metric by itself is useful for comparing lexicons of **fixed size** and for **a specific task**. Therefore, comparing the $I(L)$ of the meta-review with the one from peer reviews is an uneven comparison. The reason for this is that, as we have observed, classifying meta-reviews is a much simpler problem. This leads to lexicons that perform much better and therefore, having higher $I(L)$ values even when words related to the confounders persist in the lexicon.

An important observation is that the deconfounded lexicons are less predictive of the final outcome compared to the non-causal lexicons, when not combined with the Confounder. For instance, in the case of Peer-Review GRU model, the F1-Score of $L(T)$ drops from 60% to 59% with Deep Residualisation. On the contrary, when the deconfounded lexicons are coupled with the confounders, the logistic regression performs better compared to the non-causal lexicon. For the Peer-Review GRU model, the F1-Score of $L(T), C$ increases from 60% to 64% with Deep Residualisation. The intuition behind this is that the initial lexicon had some words related to the confounders that were important for the final prediction. The DR method does not attend to those words because they are related to the confounders. Instead, they are replaced with words, which may be less predictive of the final outcome, yet uncorrelated to the confounders.

Ultimately, DR+LIME and DR+BoW are less successful than DR+ATTN in generating a deconfounded lexicon for the peer reviews since the lexicon generated from DR+ATTN achieves a superior Informativeness Coefficient. Still, our DR+LIME technique manages to produce a more informative lexicon than the DR+BoW model and the LRC baseline.

5.4 Discussion on our Assumptions

Causal inference is limited to making several assumptions and providing causal guarantees for as long as those assumptions hold (Molnar et al., 2020; Pryzant et al., 2018a; Veitch et al., 2019). We assume that:

1. The subject of a paper is sufficient to identify the causal effect, i.e., there is no unobserved

GRU (Non-Causal)	GRU (DR+ATTN)	SciBERT (Non-Causal)	SciBERT(DR+LIME)
not, attempt, comprehensive, interfere, variational, entirely, systematically, perfect, sure, lda, formalize, tolerant, valid, belief, gan, imposible, adagrad, semantically, aware, clear, carefully, j., and/or, arguably, fully, offer, vs., confident, message, object, receive, probably, multimodality, strictly, directly, twitter, join, e.g., observe, complete, explain, novelty, dualnets, convince, handle, nintendo, theorem, satisfy, ready, demand	identical, sure, premise, carefully, heavy, specify, tease, interfere, necessarily, adagrad, attempt, novelty, fully, repeat, systematically, not, gray, theoretically, anymore, role, belief, recommendation, almost, consistently, primary, subtraction, good, satisfactory, background, write, compute, easy, teach, enough, convince, except, ready, explain, usage, sell, unfortunately, arguably, yet, especially, elegantly, sufficiently, handle, cdl, setup, reception	interspeech, zhang, prons, dismissed, third, p6, confusing, resulting, geometry, unconvincing, honestly, not, community, cannot, superficial, readership, big, suggestion, dialogue, revisit, bag, analysed, icml, taken, typo, submitted, energy, nice, spirit, competitive, per, highlighted, multimodality, far, 04562, lack, preprint, dcgan, conduct, word2vec, wen, gan, rejected, start, towards, multiplication, generalisation, auxiliary, parametric, enjoyed	interspeech, dismissed, prons, geometry, p6, submitted, unfair, unconvincing, readership, honestly, not, spirit, confusing, analysed, bag, unable, cannot, rejected, insightful, multiplication, highlighted, enjoyed, misleading, disagree, dialogue,mu_, wen, lack, nice, multimodality, welcome, conduct, recommender, encourage, dualnets, thanks, cdl, preprint, 04562, enjoy, revisit, community, appreciate, principled, medical, coding, drnn, accompanying, factorization, jmlr

Table 3: Top 50 salient words for Individual Peer Review Classification. We present class-agnostic explanations both for some of the non-causal models (Section 4) and for some DR models (Section 5).

Model Type	Model	$I(L)$	Logistic Regression Macro F1-Score		
			Only Text $L(T)$	Only Conf. C	Text and Conf. $L(T), C$
Meta Reviews	LR Baseline	0.1547	0.68	0.50	0.71
	LRC Baseline	0.1906	0.69	0.50	0.74
	GRU	0.1862	0.71	0.50	0.74
	DR+ATTN (GRU)	**0.2113**	0.65	0.50	**0.75**
Individual Peer Reviews	LR Baseline	0.0058	0.45	0.55	0.56
	LRC Baseline	0.0163	0.46	0.55	0.58
	GRU	0.0252	0.60	0.55	0.60
	DR+ATTN (GRU)	**0.0907**	0.59	0.55	**0.64**
	SciBERT	0.0115	0.58	0.55	0.58
	DR+LIME (SciBERT)	**0.0301**	0.56	0.55	**0.60**
	BoW	0.0039	0.43	0.55	0.55
	DR+BoW	**0.0128**	0.43	0.55	**0.56**

Table 4: Performance of the Deconfounded lexicons. All lexicons have a size of 50 words. For each type of classifier, we present the Informativeness Coefficient both of the non-causal lexicon and of the deconfounded lexicon, generated through Deep Residualisation (DR).

confounding. This is a standard causality assumption, albeit a strong one (Veitch et al., 2019). This assumption essentially means that there should not be other external confounders. The first way to attack this assumption is to argue that a reviewer could have biases that affect their decisions. Those biases can act as external confounders. For instance, a famous author may receive preferential treatment from some reviewers. Indeed, in "single-blind" peer reviews, as was the case for ICLR 2017, it has been observed that reviewers tend to accept more papers from top universities and companies (Tomkins et al., 2017), compared to "double-blind" peer reviews, where the author is anonymous during the review process. However, it is practically impossible to adjust for every potential source of reviewer bias. Another way to attack this assumption is by arguing that the writing quality, for instance, could be another confounder. Nevertheless, when a paper lacks clarity, we can assume that the reviewer will point out the poor writing quality in the review. In that case, the review text accounts for the writing quality and therefore, it is not an external confounder.

2. The abstract of the paper can act as a proxy to its subject. Indeed, an abstract should reflect the main topic of the paper. In that setting, the abstract can be thought of as a "noisy realisation" of the subject (Veitch et al., 2019).

3. The embedding method is able to extract information relevant to the subject of the paper. The use of state-of-the-art language models justifies this assumption (Veitch et al., 2019). In this case, we are using SciBERT embeddings, which achieve state-of-the-art performance on scientific NLP tasks (Beltagy et al., 2019). We could further bolster this assumption by fine-tuning our SciBERT model. However, this would require a larger dataset of peer

reviews to avoid overfitting.

4. The outcome $Y(t)$ is generated from the following data generated process: $Y(t) = f_c(t) + \epsilon$, where t is the text and c the fixed observed confounders. This is a standard assumption in observational studies, and along with the assumption that there is no unobserved confounding, they guarantee that the Informativeness Coefficient measures the causal effect of the full text T on the outcome Y (Pryzant et al., 2018b).

5.5 Lexicons Inspection

As long as our assumptions hold, we can causally interpret the explanations of our models. Since we value the degree with which DR has managed to deconfound our lexicon, the explanations for the peer reviews should be extracted from DR+ATTN, which achieves the greater $I(L)$ values.

In Table 3, we can observe a comparison of the top 50 salient words between the causal and non-causal models for the peer reviews. In the first place, we can observe that in the top 50 explanations of DR+ATTN, technical words, like "generative", "variational" and "lda" disappear in the explanations of the causal model. This occurrence fits our expectations: the technical words are generally related to the confounder (the subject of a paper) and should not appear in the explanation. As far as the BERT-based causal model is concerned, its lexicon still has plenty of technical terms. This is because the DR+LIME method is less successful in removing the confounding effect.

For the meta-reviews, we notice the prevalence of words like "enjoyed", "accepted", "mixed", "positively" and "recommend". Those words are, for the most part, related with to the opinions of the peer reviewers (e.g. "The reviewers all **enjoyed** reading this paper"). Furthermore, a logistic regressor fitted on the average recommendation score of the reviewers is able to predict the final outcome of the submission with an accuracy of 95%. All this evidence suggests that the ICLR 2017 committee chair for the vast majority of papers followed the recommendation of the reviewers.

In the case of peer reviews, highly ranked in our explanations, we can observe the word "novelty". We can deduce that the Originality of a paper is indeed essential for ICLR 2017. In the top 50 words we observe words such as "explain", "carefully", "systematically", "consistently", "convince" and "sufficiently". Those words indicate the importance of the Soundness of the scientific approach and how convincingly the claims are presented. Moreover, in the top 100 words, there words that reflect the Clarity and the quality of writing: "elegantly", "polish", "readable", "clear". We can, also, detect words like "comprehensive" and "extensive", which reflect the Substance (and completeness) of a paper. Lastly, various words exist that potentially reflect other aspects of a paper, such as Meaningful Comparison ("compare"), Impact ("contribute") and Appropriateness ("appropriate").

Therefore, if our assumptions hold, we can claim that certain characteristics of a paper, such as Originality, Soundness and Clarity, lead the reviewers to recommend that a scientific paper is accepted for publication.

6 Conclusions

With peer-reviewing as a use case, we have proposed a framework for interpreting the decisions of text classifiers causally, by using natural language to adjust for confounding effects. Our methodology succeeds in extracting deconfounded lexicons that exclude terms related to the paper subject (i.e., the confounders), a task where established, non-causal global explanation approaches fail. Prior work was limited to either nominal confounders (Pryzant et al., 2017, 2018a,b) or binary treatment variables (Veitch et al., 2019). Furthermore, we have extended the Deep Residualisation (Pryzant et al., 2018a) to allow for black-box models.

As is standard in causal inference (Molnar et al., 2020), the causal guarantees of the deconfounded lexicon we generate is limited by certain assumptions. Future work could explore strengthening our assumptions through fine-tuning of the Transformer Language Model representation of the confounder text. Furthermore, fine-tuning the Encoder of DR+LIME may lead to producing more informative lexicons. To avoid overfitting in this fine-tuning process, we intend to apply our methodology on larger datasets that the one used in this paper for peer reviews.

References

André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. 2010. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347.

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew E. Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. 2018. Construction of the literature graph in semantic scholar. *CoRR*, abs/1805.02262.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. Scibert: Pretrained contextualized embeddings for scientific text. *CoRR*, abs/1903.10676.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Christian Fong and Justin Grimmer. 2016. Discovery of treatments from text corpora. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1600–1609, Berlin, Germany. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. 2019. Argument mining for understanding peer reviews. *CoRR*, abs/1903.10104.

Mark Ibrahim, Melissa Louie, Ceena Modarres, and John W. Paisley. 2019. Global explanations of neural networks: Mapping the landscape of predictions. *CoRR*, abs/1902.02384.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard H. Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (peerread): Collection, insights and NLP applications. *CoRR*, abs/1804.09635.

Katherine A. Keith, David Jensen, and Brendan O'Connor. 2020. Text and causal inference: A review of using text to remove confounding from causal estimates. *CoRR*, abs/2005.00649.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2017. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav).

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Richard L Kravitz, Peter Franks, Mitchell D Feldman, Martha Gerrity, Cindy Byrne, and William M Tierney. 2010. Editorial peer reviewers' recommendations at a general medical journal: are they reliable and do editors care? *PLoS One*, 5(4):e10072.

John Langford and Mark Guzdial. 2015. The arbitrariness of reviews, and advice for school administrators. *Commun. ACM*, 58(4):12–13.

Zachary C. Lipton. 2016. The mythos of model interpretability.

Scott M. Lundberg, Gabriel G. Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2019. Explainable AI for trees: From local explanations to global understanding. *CoRR*, abs/1905.04610.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

R McNamee. 2005. Regression modelling and other methods to control confounding. *Occupational and Environmental Medicine*, 62(7):500–506.

Christoph Molnar. 2019. *Interpretable Machine Learning*. `https://christophm.github.io/interpretable-ml-book/`.

Christoph Molnar, Gunnar König, Julia Herbinger, Timo Freiesleben, Susanne Dandl, Christian A. Scholbeck, Giuseppe Casalicchio, Moritz Grosse-Wentrup, and Bernd Bischl. 2020. Pitfalls to avoid when interpreting machine learning models.

W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080.

Jonas Peters. 2017. Elements of causal inference.

Reid Pryzant, Sugato Basu, and Kazoo Sone. 2018a. Interpretable neural architectures for attributing an ad's performance to its writing style. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 125–135, Brussels, Belgium. Association for Computational Linguistics.

Reid Pryzant, Youngjoo Chung, and Dan Jurafsky. 2017. Predicting sales from the language of product descriptions. In *eCOM@SIGIR*.

Reid Pryzant, Kelly Shen, Dan Jurafsky, and Stefan Wagner. 2018b. Deconfounded lexicon induction for interpretable social science. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1615–1625, New Orleans, Louisiana. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Margaret E. Roberts, Brandon M. Stewart, and Richard A. Nielsen. 2020. Adjusting for confounding with text matching. *American Journal of Political Science*, 64(4):887–903.

Cynthia Rudin. 2018. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.

Koustuv Saha, Benjamin Sugar, John Torous, Bruno Abrahao, Emre Kıcıman, and Munmun De Choudhury. 2019. A social media study on the effects of psychiatric medication use. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):440–451.

Bernhard Schölkopf. 2019. Causality for machine learning.

Lukas Stappen, Georgios Rizos, Madina Hasan, Thomas Hain, and Björn W. Schuller. 2020. Uncertainty-Aware Machine Support for Paper Reviewing on the Interspeech 2019 Submission Corpus. In *Proc. Interspeech 2020*, pages 1808–1812.

Andrew Tomkins, Min Zhang, and William D. Heavlin. 2017. Reviewer bias in single- versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 114(48):12708–12713.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Victor Veitch, Dhanya Sridhar, and David M. Blei. 2019. Using text embeddings for causal inference. *CoRR*, abs/1905.12741.

Ke Wang and Xiaojun Wan. 2018. Sentiment analysis of peer review texts for scholarly papers. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 175–184, New York, NY, USA. Association for Computing Machinery.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation.

Partition	#Papers	Accepted / Rejected
Training Set	349	139 / 210
Validation Set	40	18 / 22
Test Set	38	15 / 23
Total	427	172 / 255

Table 5: Description of ICLR 2017 section of PeerRead.

A Dataset statistics

Dataset statistics are summarised in Table 5.

B Implementation Details

In the section, we provide implementation details such as the number of neurons in our networks and the training hyperparameters. All hidden layers of our neural networks have a ReLU activation function, and the output layers have a Sigmoid activation function.

Let us first focus on the non-causal `SciBERT` [CLS] approach. For the Final Decision Peer Review Classifier, we are using two attention layers and therefore the concatenated fused embeddings $768 * 2 = 1536$ dimensions. The first layer of the MLP has 1536 input neurons and 128 output neurons. The second layer has 128 input neurons and 64 output neurons. The third (and last) layer has 64 input neurons and 1 output neuron. The probability of the Dropout for each layer is 20%. The Individual Decision Peer Review Classifier and the Meta-Review Classifier have essentially the same MLP as the Final Decision Peer Review Classifier but with 768 input neurons. For all `SciBERT` [CLS] classifiers, we are using Binary Cross Entropy as the loss function and an Adam optimiser (Kingma and Ba, 2014) with a learning rate of 1×10^{-4}. However, we are training our classifiers for a different amount of epochs and with different batch sizes. For the Final Decision Peer Review Classifier, we are training for 500 epochs with a batch size of 100. The Meta-Review Classifier is trained for the 100 epochs, with a batch size of 100. Finally, the Individual Peer Review Classifier has 220 training epochs and a batch size of 300 samples.

Focusing on the parameters of the non-causal GRU classifiers, the GRU cell has 768 input neurons (the size of the BERT embedding) and 30 output neurons. The first layer of the MLP has 30 input neurons and 16 output neurons. The last layer has 16 input neurons and 1 output neuron. The probability of the Dropout for each layer is

20%. Similarly with before, we are using a Binary Cross Entropy loss function with an Adam optimiser. The Individual Peer Review classifier has a learning rate of 5×10^{-4} and is trained for 80 training epochs with a batch size of 500 samples. The Meta-Review classifier has a learning rate of 8×10^{-5}, is trained for 70 epochs and has a batch size of 30 samples.

Focusing on the causal DR models, all models have the same architecture for the MLP that produces the intermediate predictions $\hat{Y}'$. The first layer has 768 input neurons, 10 output neurons. The second layer has 10 input neurons, 5 output neurons. The last layer is going to have 5 input neurons, 1 output neuron.

In the case of the DR+LIME causal model, the MLP that produces the embedding e has 2 layers. The first one has 768 input neurons and 300 output neurons. The second layer has 300 input neurons and 100 output neurons. The embedding e is constructed as a 100 dimensional vector. The MLP that predicts the final acceptance decision has 3 layers. The first layer has 101 input neuron, 64 output neurons. The second one has 64 input neurons and 16 output neuron. The last layer has 16 input neurons and 1 output neuron. An Adam optimiser is used, with a learning rate of 1×10^{-4}. The model is trained for 125 epochs, with a batch size of 300.

As far the DR+LIME models are concerned, in the case of Individual Peer Review Classification, the embedding e is a 30 dimensional vector. The MLP that predicts the final acceptance decision has 2 layers. The first layer has 31 input neurons, 16 output neurons. The last layer has 16 input neurons and 1 output neuron. An Adam optimiser is used, with a learning rate of 5×10^{-4}. The model is trained for 70 epochs, with a batch size of 250. In the case of meta-review classification, the embedding e is a 64 dimensional vector. Similarly with before, The MLP that predicts the final acceptance decision has 2 layers. The first layer has 65 input neurons, 32 output neurons. The last layer has 12 input neurons and 1 output neuron. Lastly, an Adam optimiser is used, with a learning rate of 1×10^{-4}. The model is trained for 80 epochs, with a batch size of 30.

For the DR+BoW model, the MLP that predicts the final acceptance decision has a single layer with 2 input neurons and 1 output neuron. An Adam optimiser is used, with a learning rate of 8×10^{-4}. The model is trained for 200 epochs, with a batch size of 500.

B.1 Preprocessing

B.1.1 SciBERT Model Preprocessing

For the SciBERT models (Beltagy et al., 2019), we perform standard BERT preprocessing, which includes tokenisation and lowercasing. Additionally, we truncate the end of the reviews that exceed 512 input tokens, which is the maximum supported by BERT (Devlin et al., 2018). The mean length of an ICLR 2017 peer review is 346 input tokens, with a standard deviation of 213 (Kang et al., 2018). For meta-reviews, the mean length drops to 93 and the standard deviation to 77.

Truncating reviews will affect the performance of our models. However, that effect is limited to an extent. Namely, only 10.5% of the peer reviews have a length of more than 512 tokens. From that 10.5%, around half of those reviews have less than 15% of their tokens truncated. Therefore, for those reviews, the effect of the truncation is not extremely adverse. Lastly, 0.5% of all peer reviews have more than half of their tokens truncated which could have a substantial impact in the classification of those few reviews. A histogram with the lengths of peer reviews can be seen in Figure 3.

For meta-reviews, the effect of truncation is negligible since only 0.05% have more than 512 input tokens. The histogram with the lengths of meta-reviews can be seen in Figure 4.

B.1.2 GRU Model Preprocessing

For GRU models, apart from tokenisation, we perform punctuation removal, lemmatisation and lowercasing.

C Confusion Matrix

A summary of the confusion matrices can be seen in Figure 6.

D Explanations

In this Appendix, we present tables with the explanations we generate with our various models. In Table 7, we provide the top 50 important words for every non-causal model. In Table 8, we present the top 50 non-causal and causal words from meta-reviews. Table 9 contains the top 50 words for causal and non-causal Individual Peer Review Classifiers. Lastly, Table 10 provides the top salient n-grams for causal and non-causal Bag-of-Words models.

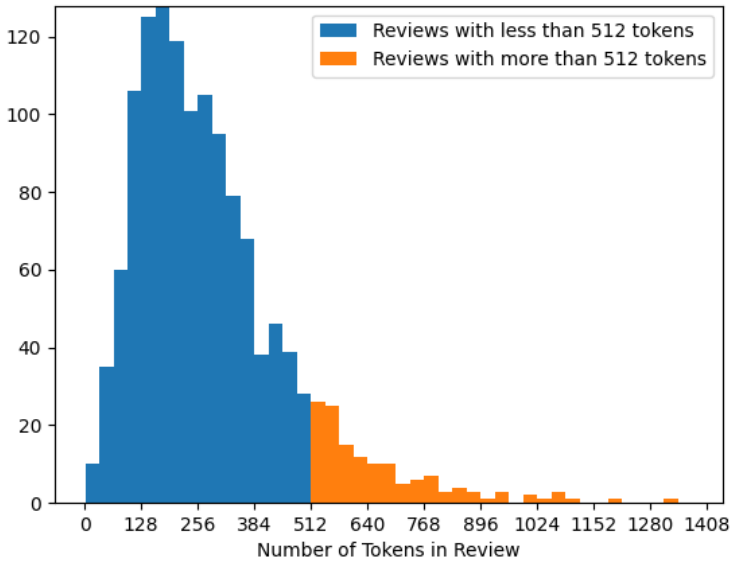

Figure 3: Histogram of the number of input tokens per peer review.

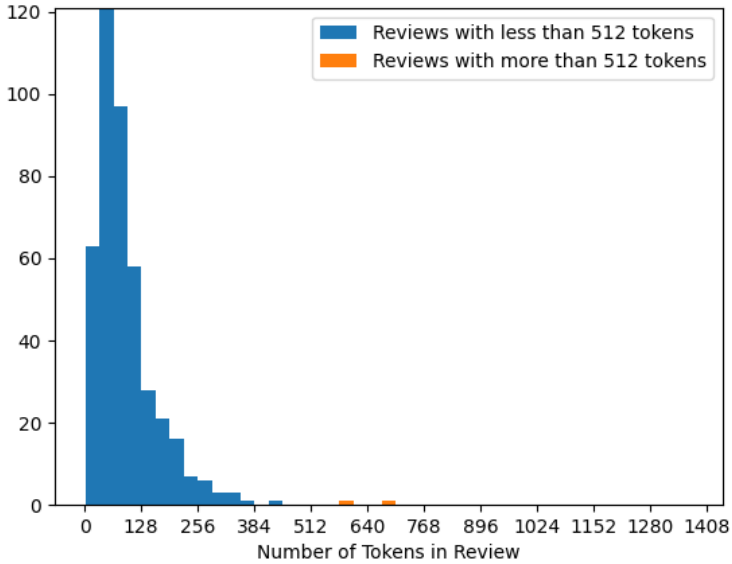

Figure 4: Histogram of the number of input tokens per meta-review.

		Predicted					
		Meta-Review SciBERT		Meta-Review GRU		Peer Review SciBERT	
		Accepted	Rejected	Accepted	Rejected	Accepted	Rejected
Actual	Accepted	13	2	14	1	12	3
	Rejected	2	21	4	19	2	21

Table 6: Final Decision Classifier Confusion Matrices on the test set of PeerRead (Kang et al., 2018)

BERT - Meta-Review	**GRU - Meta-Review**	**BERT - Individual Peer Review**	**GRU - Individual Peer Review**	**BERT - Final Decision Peer Review**
enjoyed	positively	interspeech	not	kolter
cnns	accept	zhang	attempt	aaai
resubmit	suspiciously	prons	comprehensive	not
community	wrong	dismissed	interfere	third
limited	yet	third	variational	efficiently
experimentation	perhaps	p6	entirely	untrue
3	mix	confusing	systematically	cifar10
workshop	limit	resulting	perfect	venue
poster	guidance	geometry	sure	recovered
topic	read	unconvincing	lda	huffman
variational	impressive	honestly	formalize	riemannian
limit	surprise	not	tolerant	cifar
although	enough	community	valid	arxiv
not	incremental	cannot	belief	accompanying
sound	sufficiently	superficial	gan	extended
none	timely	readership	imposible	convex
accepted	journal	big	adagrad	github
dropout	post	suggestion	semantically	trace
imagenet	contribute	dialogue	aware	eigenspectra
agent	ready	revisit	clear	encourage
justify	not	bag	carefully	code
generative	variational	analysed	j.	asynchronous
advantage	nicely	icml	and/or	crucial
unfortunately	receive	taken	arguably	auxiliary
video	submit	typo	fully	investigated
review	relevant	submitted	offer	iclrw2016
solver	much	energy	vs.	hyperparameters
three	high	nice	confident	discus
task	weak	spirit	message	dcgan
use	good	competitive	object	interesting
confident	write	per	receive	nice
generalisation	whilst	highlighted	probably	game
supervised	into	multimodality	multimodality	demonstration
promise	below	far	strictly	lample
raised	benefit	04562	directly	the
understanding	rebuttal	lack	twitter	significance
timely	generative	preprint	join	seq2seq
reject	strengthen	dcgan	e.g.	community
unsatisfactory	meet	conduct	observe	perspective
yet	variety	word2vec	complete	paper
mixed	give	wen	explain	thanks
shot	hide	gan	novelty	investigates
propose	recommend	rejected	dualnets	insightful
how	badly	start	convince	uploaded
rejection	long	towards	handle	narrow
thus	clear	multiplication	nintendo	match
salimans	convince	generalisation	theorem	seed
provided	gain	auxiliary	satisfy	tsp
goal	serious	parametric	ready	adversarial
reservation	maintain	enjoyed	demand	isn

Table 7: Top 50 salient words for non-causal models.

GRU (Non-Causal)	GRU (DR+ATTN)
positively	nicely
accept	overcomplicated
suspiciously	positively
wrong	yet
yet	rebuttal
perhaps	ready
mix	sufficiently
limit	properly
guidance	acceptance
read	surprise
impressive	usual
surprise	into
enough	describe
incremental	finding
sufficiently	organize
timely	execution
journal	insight
post	receive
contribute	criterion
ready	overall
not	good
variational	closely
nicely	wise
receive	suspiciously
submit	well
relevant	badly
much	write
high	pair
weak	explain
good	not
write	simple
whilst	weak
into	asset
below	great
benefit	out
rebuttal	execute
generative	helpful
strengthen	enjoy
meet	conditional
variety	mix
give	entirely
hide	contribution
recommend	world
badly	demonstrate
long	accept
clear	assume
convince	post
gain	publication
serious	effective
maintain	clearly

Table 8: Top 50 salient words for meta-reviews. The first column contains the explanations of the non-causal models. The second column contains the deconfounded lexicon.

BERT (Non-Causal)	BERT (DR+LIME)	GRU (Non-Causal)	GRU (DR+ATTN)
interspeech	interspeech	not	identical
zhang	dismissed	attempt	sure
prons	prons	comprehensive	premise
dismissed	geometry	interfere	carefully
third	p6	variational	heavy
p6	submitted	entirely	specify
confusing	unfair	systematically	tease
resulting	unconvincing	perfect	interfere
geometry	readership	sure	necessarily
unconvincing	honestly	lda	adagrad
honestly	not	formalize	attempt
not	spirit	tolerant	novelty
community	confusing	valid	fully
cannot	analysed	belief	repeat
superficial	bag	gan	systematically
readership	unable	imposible	not
big	cannot	adagrad	gray
suggestion	rejected	semantically	theoretically
dialogue	insightful	aware	anymore
revisit	multiplication	clear	role
bag	highlighted	carefully	belief
analysed	enjoyed	j.	recommendation
icml	misleading	and/or	almost
taken	disagree	arguably	consistently
typo	dialogue	fully	primary
submitted	mu_	offer	subtraction
energy	wen	vs.	good
nice	lack	confident	satisfactory
spirit	nice	message	background
competitive	multimodality	object	write
per	welcome	receive	compute
highlighted	conduct	probably	easy
multimodality	recommender	multimodality	teach
far	encourage	strictly	enough
04562	dualnets	directly	convince
lack	thanks	twitter	except
preprint	cdl	join	ready
dcgan	preprint	e.g.	explain
conduct	04562	observe	usage
word2vec	enjoy	complete	sell
wen	revisit	explain	unfortunately
gan	community	novelty	arguably
rejected	appreciate	dualnets	yet
start	principled	convince	especially
towards	medical	handle	elegantly
multiplication	coding	nintendo	sufficiently
generalisation	drnn	theorem	handle
auxiliary	accompanying	satisfy	cdl
parametric	factorization	ready	setup
enjoyed	jmlr	demand	resception

Table 9: Top 50 salient words for the task of Individual Peer Classification. In order to make comparisons, we present both causal and non-causal models.

BoW (Non-Causal)	BoW (DR+BoW)
(not, enough)	(not, enough)
(much, be)	(much, be)
(literature, and)	(not, compare)
(need, a)	(evidence, to)
(evidence, to)	(be, insufficient)
(be, not, enough)	(comparison, with, other)
(not, compare)	(be, mention)
(on, use)	(literature, and)
(comparison, with, other)	(need, a)
(doe, not, provide)	complementary
(claim, that, this)	(be, not, enough)
(network, weight)	(method, need)
(do, not, compare)	(would, require)
(work, need)	(doe, not, provide)
gans	(on, use)
(layer, this)	(to, compare, the)
(be, insufficient)	(have, see)
(it, look, like)	(comparison, with, the)
(push, the)	(network, weight)
(work, so)	(do, not, compare)
(to, compare, the)	(sure, if, the)
(not, compare, to)	(no, quantitative)
(not, surprise)	(no, new)
(and, compare, to)	(experiment, there, be)
(have, see)	(on, the, previous)
(very, hard)	(improve, version)
(of, the, dataset)	(not, compare, to)
(be, mention)	(adversarial, train, to)
(should, be, mention)	(dataset, which, be)
(not, find, the)	(limit, in)
(not, ready)	(experiment, there)
(novelty, and)	(layer, this)
(no, quantitative)	(work, so)
(the, theoretical, analysis)	(be, write)
(experiment, there)	(the, basis)
(no, new)	(claim, that, this)
(to, control)	(not, ready)
complementary	(work, need)
(would, require)	(the, theoretical, analysis)
(enough, detail)	incomplete
(train, neural, network)	(reasonably, good)
(of, video)	(it, hard, to)
(the, combination, of)	(a, system)
incomplete	(and, compare, to)
(the, claim, about)	gans
(they, introduce)	(they, introduce)
(comparison, on)	(result, be, provide)
(limit, in)	(not, find, the)
push	(should, be, mention)
(cover, a)	(push, the)

Table 10: Top 50 salient n-grams for peer reviews and top 50 deconfounded n-grams.