APCS: Towards Argument based Pros and Cons Summarization of Peer Reviews

Sandeep Kumar†, Tirthankar Ghosal*, Asif Ekbal†
†Indian Institute of Technology Patna, India
∗National Center for Computational Sciences, Oak Ridge National Laboratory, USA
†(sandeep_2121cs29,asif)@iitp.ac.in
∗ghosalt@ornl.gov

Abstract

Peer review is an evaluation process where experts in a particular field assess the quality and credibility of a research paper or manuscript prior to its publication. Utilizing Artificial Intelligence (AI) in the peer review process has the potential to enhance the review process by providing more objective, efficient and accurate evaluations. Summarizing the pros and cons of peer reviews will be valuable for editors/area chairs to provide constructive feedback to authors, make informed decisions about manuscript publication and identify potential issues in the field. It will also assist them in understanding which areas of their work need improvement and which do not. In this research, we propose a novel architecture that uses a supervised method to generate generic pros and cons summaries to assist editors and authors in analyzing the feedback from peer reviews. Additionally, we propose an unsupervised method for generating aspect-based pros and cons summaries. Our proposed method achieves an average Rouge-1 F1 Score of 31.61 in generating generic pros and cons summaries and 32.62 in generating aspect-based summaries.

1 Introduction

Peer review, a process in which experts in a specific field assess the quality of research work, is a vital aspect of scientific discovery. It is well-known that peer reviews are controversial due to their quality, randomness, bias, and inconsistencies (Bornmann and Daniel, 2010). Additionally, there have been concerns about alleged reviewer bias in "single-blind" peer reviews (Tomkins et al., 2017) and arbitrariness between different reviewer groups (Langford and Guzdial, 2015). Despite these criticisms, within the scientific community, peer review is considered as an essential component of the academic writing process as it helps ensure that the papers published in scientific journals are of high quality and based on accurate experimentation. However, despite its significance, there is a lack of analysis and evaluation of the content and structure of reviews and their quality. According to a study by Kovanis et al. (Kovanis et al., 2016), approximately 63.4 million hours were spent on peer reviews in 2015 alone. The rapid increase in the number of publications in scientific fields motivates the development of automatic summarization tools for scientific articles. The number of scientific articles published per year has been growing at a rate of about 8% per year since the mid-17th century (Kovanis et al., 2016). The number of scientific papers indexed in the Web of Science database has been increasing at a rate of about 3% per year since the 1970s.

Investigating the inner workings of the peer review system can be challenging due to the need to protect publishers' privacy and intellectual property rights. However, OpenReview[1] provides a way to examine how the process is evolving in some areas, such as how authors are given opportunities to respond to feedback and how communication between authors and reviewers is being strengthened.

Argument mining in peer-review text is an important tool in the scientific publication process as it enables the automated analysis and extraction of key claims, evidence, and reasoning presented in a manuscript. This improves the efficiency, consistency, and fairness of the review process, detects potential biases, and assists authors in identifying areas for improvement, ultimately leading to a higher-quality manuscript and aiding in the advancement of scientific knowledge. Argument Mining can be used to efficiently extract the most relevant parts from reviews, which are paramount for the publication decision. Fromm et al. (Fromm et al., 2020) propose a simple argumentation scheme that distinguishes between non-arguments, supporting arguments, and attacking

[1] https://openreview.net/

Proceedings of the 2nd Workshop on Information Extraction from Scientific Publications, pages 117–129
Nov 1, 2023. ©2023 Association for Computational Linguistics

	Summary
Pros:	The paper introduces a novel approach for sentence representation by using multiple attentional vectors to extract multiple representations for a sentence. The authors have demonstrated consistent gains across three different tasks, providing evidence of the effectiveness of the model. The paper is reasonably clear, with no major technical issues, and the new model lends itself to more informative visualizations than could be obtained otherwise. The model also beats reasonable baselines on three datasets. The architecture is interesting and can be used within larger text understanding models. The approach is different from prior work, which is a positive aspect of the paper.
Cons:	a lack of analysis on the 2D representations, concerns about the value of r when applied to short sentences, a need for performance evaluations on dev sets or learning curves, and a lack of transparency in reporting model sizes. The paper also has a problem in its presentation, with no training objective defined, and there is a lack of appropriate addressing of prior work. The visualizations provided do not offer compelling evidence for the use of multiple attention vectors, and further experiments are needed to demonstrate the effectiveness of the 2D structure of the embedding matrix. Overall, there is a lack of convincing evidence that the 2D structure of the embedding matrix provides any meaningful advantage over similar attentive embedding models.

Table 1: Pros and Cons summary output of paper (ICLR 2017); https://openreview.net/forum?id=BJC_jUqxe

Aspects	Summary
Substance	Pros: The paper introduces a novel approach for sentence representation using 2D structure of embeddings, which produces more informative visualizations and beats reasonable baselines on three datasets. Cons: the reviewer would like to see more analysis on the 2D representations in order to be convinced of its effectiveness ablation studies?
Clarity	Pros: The paper is reasonably clear and there are no major technical issues. Cons: there are issues with the penalization term section and the paper's focus on unsupervised learning in the abstract, introduction and related work sections, and with the lack of clear definition of the training objective.
Meaningful Comparison	Cons: There is a substantial amount of prior work which the authors do not appropriately address , some of which is listed in previous comments .
Originality	Pros: the main innovation of this paper is the 2D structure of the embedding matrix Cons: 2D structure of the embedding matrix is not clearly shown to provide significant advantages over similar attentive embedding models already present in the literature.
No-aspect	Pros: This paper presents a method for sentence representation using a 2D matrix and self-attentive mechanism on LSTM encoder. It produces heat-map visualizations and good performance on downstream tasks. The model extracts matrix-valued sentence representation and could be used for tasks beyond NLP. The authors have shown consistent gains across multiple datasets. Cons: Some important experiments are missing, visualizations lack support for multiple attention vectors, main claims require more experimentation, unclear usage and conversion of embedding for downstream tasks, better model structure explanation needed, no comparison with similar works, minor issues like typos present.

Table 2: Aspect wise Pros and Cons summary output

arguments (NON/PRO/CON) as outlined in (Stab et al., 2018). This scheme can also be interpreted as a simplified version of the claim-premise model, where if there is a single claim, "The paper should be accepted," and arguments that either support or attack this claim.

An editor or chair writes a meta-review evaluating and summarizing the strengths and weaknesses of a peer review process as it pertains to a specific research or manuscript. Classification of meta-review is important because it allows readers to evaluate the quality and reliability of the research presented in the text and make informed decisions about its validity and usefulness. Additionally, it is important for researchers as it allows them to identify areas of improvement in their own research and writing process. Furthermore, it is essential for editors as it enables them to provide constructive feedback to authors, make informed decisions about the publication of a manuscript, and identify potential issues in the field. Thus, meta-review and its classification play a vital role in the scientific publication process. MReD dataset (Shen et al., 2022) consists of 7,089 meta-reviews and all its 45k meta-review sentences. Each sentence in a meta-review is classified into one of the 9 pre-defined intent categories: abstract, strength, weakness, rating summary, area chair (AC) disagreement, rebuttal process, suggestion, decision, and miscellaneous(misc).

Summarizing the pros and cons of a peer review text is crucial as it provides readers with a comprehensive understanding of the strengths and weaknesses of the peer review process as it pertains to a specific research or manuscript. This enables them to evaluate the quality and reliability of the research presented in the text, and make informed decisions about its validity and usefulness. Moreover, summarizing the pros and cons of a peer review text is of great importance for researchers, as it allows them to identify areas of improvement in their own research and writing process. For example, if a manuscript is rejected due to poor methodology, researchers can focus on addressing and improving that aspect of their work in future submissions. Furthermore, summarizing the pros and cons of a peer review text is essential for editors/area chair, as it enables them to provide constructive feedback and to make informed decisions about the publication of a manuscript, and identify potential issues in it, which can lead to taking appropriate steps to address them. Review text also contains aspects associated with it, such as novelty and motivation. Editors would benefit from knowing the specific pros and cons that reviewers have written about each aspect. In this research, we propose a way to generate both a generic pros and cons summary, as well as an aspect-wise pros and cons summary. This information can assist editors/area chair in quickly understanding which aspects of the paper need improvement and which do not, and can be beneficial for author as well to get a quick overview of the reviews. To demonstrate this, we present output from our proposed architecture in Table 1 and Table 2.

There exist reference summary for pro and con summary. Also the generation of human based summaries is expensive and require domain experts to summary. The meta reviewer usually mentions opinions about the submission's strengths and weakness as opinions about the submission's weaknesses. As strength mentioned in the meta review is mostly the summary of the pro argument and strength mentioned in the meta review is mostly the summary of the con argument of a paper. We used this idea and used the strength and weakness mentioned in the meta review as the reference summary.

We summarize our contributions as follows :-

- We propose an effective architecture that utilizes a supervised method for generating generic pros and cons summaries, to assist the editors and authors in analyzing peer reviews.
- We investigate the utilization of meta-reviews for this task without the availability of a reference summary for training.
- We propose a novel architecture that utilizes an unsupervised method for generating aspect-based pros and cons summaries for the same task.
- We have annotated 150 papers with aspect-based summaries to evaluate the generated aspect-based summary.

We make our code public[2].

2 Related Work

2.1 AI in Peer Reviews

The use of artificial intelligence in peer review has been garnering attention due to recent advancements in AI research. A dataset of scientific peer reviews was made available to facilitate research in this domain(Kang et al., 2018). Additionally, various studies have explored the correlation between overall recommendation scores and individual aspect scores. The CiteTracked dataset was introduced to ascertain the impact of citations from peer reviews(Plank and van Dalen, 2019). Furthermore, tools have been developed to analyze the quality, tone, and quantity of peer review comments, such as those mentioned in(Wicherts, 2016). The ASAP-Review dataset was formulated with the objective of automating scientific peer review(Yuan et al., 2021). Recently, a novel multitasking system was proposed, which leverages inter-dependency by sharing representations between two related tasks, such as aspect categorization and sentiment classification(Kumar et al., 2021). Shallow linguistic features, for instance, sentiment words, have been studied by Bornmann et al. to analyze language use in peer reviews(Bornmann et al., 2012).

2.2 Abstractive and Extractive Summarization

Extractive summarization involves creating summaries by selecting key sentences or phrases directly from the source text, retaining the original content's phrasing(Collins et al., 2017). Initially, extractive methods relied on simple statistical measures such as word frequency(Luhn, 1958b) and document location(Baxendale, 1958). As research evolved, classifiers using supervised learning identified potential summary sentences(Kupiec et al., 1995). Factors like sentence position, length, title words, and the presence of proper nouns became crucial cues(Yang et al., 2017; Nenkova et al., 2006). Modern extractive summarization predominantly employs neural models, integrating embeddings, CNNs, and RNNs(Kobayashi et al., 2015; Cheng and Lapata, 2016), and these systems often rank sentence salience before summarization(Erkan and Radev, 2004; Parveen et al., 2016).

Conversely, abstractive summarization crafts novel sentences and may use words not found in the source text(Widyassari et al., 2022). Although it offers more flexible summaries, the complexity of generating new content requires advanced natural language processing(Gambhir and Gupta, 2017). The encoder-decoder paradigm has emerged as a prominent technique in abstractive summarization(Xu et al., 2020; Lee et al., 2020; Yao et al., 2020), enabling efficient parameter optimization

[2] https://github.com/sandeep82945/Pros-Cons-Summarization-of-peer-reviews

and smoother summary generation.

2.3 Review Summarization

Several studies have explored the summarization of product reviews(Li et al., 2010; Gerani et al., 2014, 2019; Mason et al., 2016). For instance, Gerani et al.(Gerani et al., 2014) proposed an abstractive summarization system for product reviews, utilizing a template-based Natural Language Generation (NLG) framework and leveraging the discourse structure of reviews.

Aspect-based summarization involves generating focused summaries based on specific points of interest. WikiAsp(Hayashi et al., 2021), a large-scale dataset for multi-domain aspect-based summarizations, has been introduced. One study was conducted to provide insights into hotels that ratings might not fully capture by analyzing customer reviews from hotel booking websites. The topic modeling technique, Latent Dirichlet Allocation (LDA), was applied to uncover hidden information and aspects, followed by sentiment analysis on classified sentences and summarization(Akhtar et al., 2017). An interactive attention mechanism was proposed for aspect- and sentiment-aware abstractive review summarization(Yang et al., 2018). The model(Kunneman et al., 2018) incorporates representations of context, sentiment, and aspect words within reviews into the summary generation process. The authors developed three systems for generating pros and cons summaries of product reviews, which included a system based on syntactic phrases and two neural-network-based systems. These systems were evaluated in two ways: using held-out reviews with gold-standard pros and cons, and by soliciting human annotators to rate the systems' outputs in terms of relevance and completeness.

2.4 Peer Review Summarization

Peer-review summarization is a specific task that aims to automatically generate a summary of peer reviews for a particular research paper. Numerous studies have focused on this task, employing various techniques and models. Several works have built systems to generate meta-reviews from peer reviews by summarizing them.

The authors present MetaGen(Bhatia et al., 2020), a system that generates meta-reviews from peer reviews to aid the decision-making process in scientific papers and proposals. It utilizes an extractive and fine-tuned UniLM approach for crafting final abstractive meta-reviews and making acceptance/rejection decisions. A deep neural architecture was proposed for generating decision-aware meta-reviews from peer reviews(Bhatia et al., 2020). The model employs a multi-encoder transformer network for predicting the decision and generating the meta-review.

Previous studies have employed classification and regression techniques to evaluate the quality of scientific papers through analysis of peer reviews. Additionally, some research has focused on generating meta-reviews by summarizing the content of multiple reviews. To the best of our knowledge, our work is the first to summarize the argument based pros and cons of peer reviews.

3 Methodology

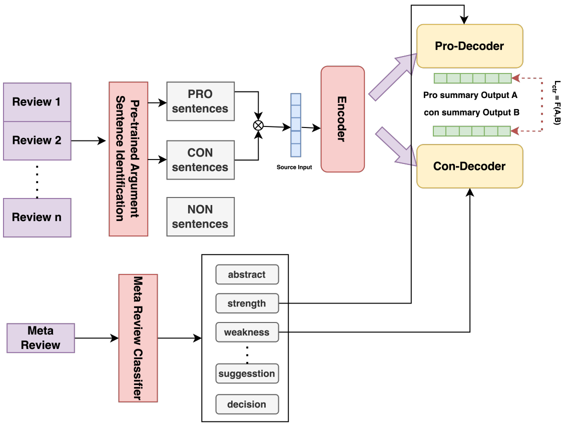

Figure 1: Our proposed architecture for generic Pros and Cons summarization

Figure 1 shows the architecture of our proposed model for generating pro and con summary.

3.1 Input Layer

Initially, we have a group or set of reviews $D = \{R_1, R_2, ..., R_n\}$, associated with a specific document or article. We merge all the reviews of a document into one comprehensive review S. Each $S = \{s_1, s_2, ..., s_m\}$, is a set of sentences, where $s_i \in S$ denotes a single sentence.

3.2 Argument Classification

Next, the set of sentences S are passed with a classifier to identify those review sentences which are argumentative. Following (Fromm et al., 2020) , we utlized a BERT large model with 340M parameters fine-tuned on the Argument Mining dataset (based on bert-large-cased) to classify the sentences into pro, con and non summary. We

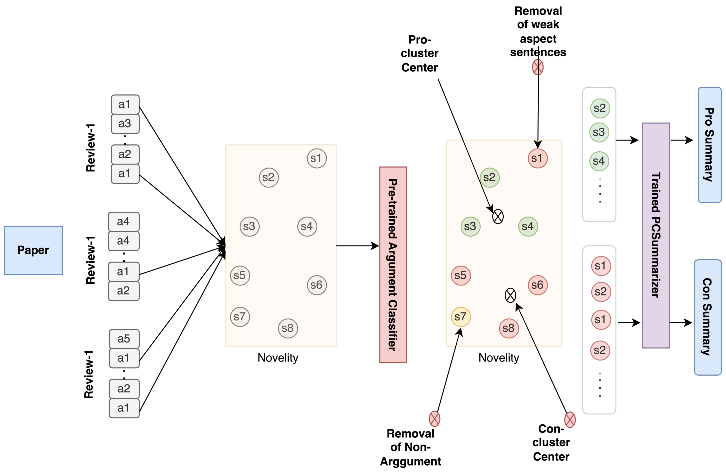

Figure 2: Our proposed aspect based pros and cons summarization architecture

reported a micro F1 score of 0.759%, which is almost the same as the original paper. For example :-

$$S_{pros} = \{s_1, s_3, ...s_{n-1}\}$$
$$S_{cons} = \{s_2, s_5, ...s_{n-3}\}$$
$$S_{nons} = \{s_4, s_6, ...s_n\}$$

Here, S_{pros} contains a set of sentences classified as pro, S_{cons} contains the sentences classified as con and S_{non} contains the sentences classified as $nons$. Non-argument sentences typically do not contain important information that is necessary for making a decision. Therefore, we discarded them.

3.3 Meta Review Classification

The reference summary of pros and cons summary is unavailable and the annotation of the summary is costly and need domain experienced experts. So we utilized the MReD dataset (Shen et al., 2022) which consists of 7,089 meta-reviews and all its 45k meta-review sentences. Each sentence in a meta-review is classified into one of the 9 pre-defined intent categories: abstract, strength, weakness, rating summary, area chair (AC) disagreement, rebuttal process, suggestion, decision, and miscellaneous(misc). We trained the RoBERTa-large + CRF with the same setting as mentioned in the paper. We hypothesize that a meta-reviewer will mention both the strengths and weaknesses of a product or research study in their summary, akin to a pro and con summary. So, we used the pre-trained model to extract the strength and weaknesses. Suppose, M is the set of review sentences in meta review. We utilize the set of sentences classified into strength $M_{strength} \in M$ and belonging to weakness $M_{weaknesses} \in M$ for training PCSummarizer, described in the next section.

3.4 PCSummarizer

Generative pre-trained models have exhibited outstanding results in the field of natural language generation, specifically in the area of text summarization (Dong et al., 2019; Lewis et al., 2020a). The adaptation of natural language processing models to specific domains, also known as domain adaptation, is a widely researched topic (Fu and Liu, 2022; III, 2009; Yu et al., 2021). Hua and Wang (2017) (Hua and Wang, 2017) were the pioneers in researching the adaptation of neural summarization models to specific domains, and it was found that these models possess the capability to select pertinent information even when trained on out-of-domain data.

In order to make the model capture the argumentative reviews (i.e. both pro and con sentences), the input text is formatted in the following way as source input for the Encoder.

Pro sentences [SEP] Con sentences

Here [SEP] is a special token.

The encoder first transforms the input into a sequence of hidden representations M.

$$h_t = ProsDecoder(M, y_{t-1}) \quad (1)$$

We initialized the ProsDecoder i.e. decoder for pros summary generation with the pre-trained Bart Large (Lewis et al., 2020a) decoder trained on CNN-daily mail. We implement the teacher forcing method on the ProsDecoder with the $M_{strength}$ to fine-tune the decoder.

$$P(y_t|y_{<t,C})^{(k)} = softmax(W_d h_t + b_d) \quad (2)$$

where h_t is the hidden representation of y_t (the t-th word in the target summary). k is the probability of generating the k-th token y_t given the previously generated $< t$ tokens and some context C.

We maximize the conditional log likelihood for a given N observation
$(C^{(i)}, Y^{(i)})_{i=1}^{N}$

$$L_{pros} = -\sum_{i=1}^{i=N}\sum_{t=1}^{t=T} logP(y_t^{(i)}|y_{<t}^{(i)}C^{(i)}) \quad (3)$$

Similarly, we define the ConDecoder :-

$$h_t = ConsDecoder(M, y_{t-1}) \quad (4)$$

The ConsDecoder (i.e. decoder for cons summary generation) is initialized with the pre-trained Bart Large decoder trained on CNN-daily mail. We implement the teacher forcing method on the ConsDecoder with the $M_{weakness}$ to train the decoder.

$$P(y_t|y_{<t,C})^{(k)} = softmax(W_d h_t + b_d) \quad (5)$$

We maximize the conditional log likelihood for a given N observation
$(C^{(i)}, Y^{(i)})_{i=1}^{N}$

$$L_{cons} = -\sum_{i=1}^{i=N}\sum_{t=1}^{t=T} logP(y_t^{(i)}|y_{<t}^{(i)}C^{(i)}) \quad (6)$$

We introduced an appropriate loss function as defined below to ensure that the similar summaries are not generated for pros and cons :-

$$L_{diss} = sim(S_{pros}, S_{cons}) \quad (7)$$

Here, sim is the similarity between the two summaries. We calculate the similarity between the two summaries by [CLS] pooling as in BERT(Devlin et al., 2019).

We employ the following loss function as our final training loss :-

$$L = L_{pros} + L_{cons} + L_{diss} \quad (8)$$

Here, we combine the MLE loss from the ProsDecoder and ConsDecoder and the dissimilarity loss while training the summarizer.

3.5 Aspect based pros and cons summarization

In this section, we describe our proposed architecture for aspect-based pros and cons summarization. Figure 2 shows the architecture of our aspect based pro and con summarization. Supervised training is not possible due to the unavailability of golden pros and cons summary for each aspect. So, we propose an unsupervised technique. Similar to the previously described input layer, the reviews are combined. The reviews are then passed to the aspect classifier. We use the already annotated dataset for our evaluation. Suppose the output after the aspect classification is S_a, where S is the set of sentences that belongs to aspect category a. The sentences belonging to each aspect S_a are passed to the argument classifier, which classifies the pre-trained argument classifier as described in Section 3.2. The output is S_a^{pros}, S_a^{cons}, S_a^{nons}. Similarly, as the non-arguments S_a^{nons} do not play much role in the decision, it is filtered out.

3.5.1 Clustering

To remove the review sentences that weakly belong to an aspect category, we produce a vector A_i for each aspect category by computing the average of sentence embeddings of the sentences belonging to that aspect. In particular, for each aspect category, we produce a vector that best represents the category, represented by the centroid. We create the sentence embedding by pre-trained BERT [CLS] pooling. We further control the selection of review sentences by filtering them based on their aspect category. In particular, we select the review sentences in S_a^{pros} and S_a^{cons}, given the aspect category a it belongs to. Our goal is to select the review sentences close to their aspect centroid. We iterate through every review sentence sen in S_a^{pros} and S_a^{cons}, we add it to the filter review set Sf_a^{pros} and Sf_a^{cons} if $cos(E_i, A_k) \leq \theta$[3]. Where E_i is the embedding of the review sentence sen.

3.5.2 Summarization

Next, we create a abstract summary of the Sf_a^{pros} and the Sf_a^{cons}. We used the model PCSummarizer trained to create pros and cons summary of the review to create the final summary for each aspect as described in Section 3.4. If review text $Sf_a^{pros/cons}$ is less than 30 words, i.e. short reviews, they don't need further summarization as they are already con-

[3] We set the threshold θ as 0.5 empirically

Conference	Number of papers	Number of reviews	Acceptance rate	avg words
ICLR 2017	427	1,304	67%	399
ICLR 2018	907	3,499	35%	403
ICLR 2019	1,419	4,332	35%	403
ICLR 2020	2,213	6,722	27%	409

Table 3: Dataset statistics

cise. Using PCSummarizer may not add any value, so in that case, we don't further summarize from it.

4 Experiments

4.1 Dataset

We use the dataset collected from OpenReview[4] by the papers (Yuan et al., 2022; Fromm et al., 2020). The dataset contains the reviews from computer-science conferences. Table 3 shows the statistics of the dataset used. For training PCSummarizer we split the dataset into 0.7, 0.1, 0.2 for training, validation and test respectively. To evaluate our aspect-based summarization method, we recruited two expert NLP annotators with a strong command of the English language. They generated summaries for 150 papers from the dataset presented in (Wicherts, 2016), which contains peer reviews classified into different aspects. The definition of these aspects is provided in Appendix Table 8.

4.2 Implementation details

For PCSummarizer, we use the BART large model pre-trained on CNN/DailyMail dataset from the hugging face library [5]. We initialized the pre-trained weights to both the decoder and the encoder before fine-tuning them. We performed hyperparameter tuning on the validation set and reported the best-performing parameters. We use a dynamic learning rate, warm up 1000 iterations, and decay afterwards. We trained the model for 10 epochs with a batch size of 4. We train all the models on a single GPU (NVIDIA A100-PCIE 40GB).

4.3 Result and Analysis

Tables 4 and 5 present the results of a comparison between the various summarization methods, including extractive methods (LexRank, TextRank, SumBasic, Luhn) and abstractive methods (Pegasus and Bart) for summarization without aspects and with aspects, respectively. The results indicate that the abstractive methods performed better than the extractive methods in terms of the ROUGE score for both the summarization tasks. The pros and cons were separately input into the extractive systems, and we report the average. Similarly, for aspect-based pros and cons summarization, we calculated the score aspect-wise for each aspect and reported the average. BERTScore (Zhang et al., 2020b) computes a similarity score between each token of a candidate sentence and that of a reference sentence, relying on contextual embeddings to calculate token similarity, as opposed to exact matches. BERTScore is mainly used in abstractive summarization, so we also report BERTScore for the abstractive baselines Pegasus and BART. Similar to the extractive summarization, we trained the pros and cons encoder and decoder architecture separately and reported the average. We found that BART performed better compared to Pegasus with 1.63 F1 BERTScore and 2.12 Rouge-1 F1 score for full reviews pros and cons and BART with 1.96 BERTScore and 0.6 Rouge-1 F1 score for aspect-based summarization. Our proposed method APCS performed better than simple BART with 0.71 BERTScore and 1.21 Rouge-1 F1 score points for full reviews and 0.75 BERTScore and 1.68 Rouge-1 F1 score for aspect-based summarization. As we used the pre-trained model for argument classification and meta review classification, we don't report those results. However the result can be found in the original paper.

4.4 Ablation Study

We analyze the effectiveness of our proposed model (APCS) by conducting an ablation study, as shown in Table 7. By comparing the results of "APCS w/o diss" in Table 7 with an improvement of 0.93 and the original BART with a distinct encoder in Table 4, it is evident that inputting the pros and cons together improves the results compared to training them separately. This is likely due to the fact that sharing an encoder allows the model to learn general features that are useful for both summarization tasks.

When we ran the model (APCS without differentiation loss), we observed that the generated summaries for the cons sometimes included information that was more appropriate for the pros. This may be due to the fact that during the annotation

[4] https://openreview.net

[5] `https://huggingface.co/`

Model	BERTScore			Rouge		
	P	R	F1	R1	R2	RL
LexRank (Erkan and Radev, 2011)	–	–	–	24.30	5.90	25.18
TextRank (Mihalcea and Tarau, 2004)	–	–	–	24.32	5.89	25.12
LSA (Ozsoy et al., 2011)	–	–	–	25.88	6.20	25.72
Luhn (Luhn, 1958a)	–	–	–	26.26	6.18	25.81
KL-Sum (Haghighi and Vanderwende, 2009)	–	–	–	27.43	6.89	25.87
Pegasus (Zhang et al., 2020a)	50.17	49.55	49.98	28.42	7.05	26.32
BART (Lewis et al., 2020b)	50.73	52.65	51.61	30.40	8.76	27.14
APCS	**51.43**	**53.43**	**52.32**	**31.61**	**9.12**	**28.80**

Table 4: Experimental results on generic pros and cons summarization; ROUGE(F1), BERTScore. Here, P→Precision, R→Recall, R1→ROUGE with unigram, R2→ROUGE-2 for bigram overlap, RL→ROUGE-L for Longest Common Subsequence

Model	BERTScore			Rouge		
	P	R	F1	R1	R2	RL
LexRank (Erkan and Radev, 2011)	–	–	–	26.30	7.86	27.17
TextRank (Mihalcea and Tarau, 2004)	–	–	–	26.29	7.81	27.10
LSA (Ozsoy et al., 2011)	–	–	–	27.82	8.16	27.71
Luhn (Luhn, 1958a)	–	–	–	28.11	8.12	27.59
KL-Sum (Haghighi and Vanderwende, 2009)	–	–	–	29.39	8.78	27.86
Pegasus (Zhang et al., 2020a)	51.67	51.64	51.60	30.29	7.05	28.31
BART (Lewis et al., 2020b)	52.71	54.62	53.56	32.41	10.74	29.11
APCS	**53.41**	**55.42**	**54.31**	**32.62**	**11.09**	**30.79**

Table 5: Experimental results on aspect-based pros and cons summarization

	w/o Aspect				Aspect based			
Model	A-Coverage	Readability	Diversity	I	A-Coverage	Readability	Diversity	I
KL-Sum(Erkan and Radev, 2011)	3.0	**8.5**	3.0	3.0	5.0	8.0	3.5	3.5
Pegasus (Zhang et al., 2020a)	4.0	4.5	3.0	3.0	5.5	4.5	3.0	3.5
BART	4.25	5.0	4.5	4.0	4.5	**5.0**	4.5	6.0
APCS	**4.5**	5.0	**5.0**	**4.25**	**7.25**	**5.0**	**5.0**	**6.25**

Table 6: Human evaluation results. Here, A-Coverage denotes Aspect coverage; I → Informativeness ; Bold text is intended to highlight the best performance.

	R-1	R-2	R-L
APCS w/o diss	31.33	9.02	28.24
APCS(aspect based) w/o clustering filter	32.03	10.78	30.12

Table 7: Ablation study of our experiments

process, reviewers/editors often use polite language when discussing cons/weakness, such as "I like the paper but..." or "The paper is written well but there are a few technical...". As a result, the ConsDecoder may have learned to include some pros information in the summary as well during the training process. We observed a slight improvement in the results when the differentiation loss was included in the model, which resulted in a better separation of the pros and cons summaries.

For aspect-based unsupervised summarization, we also removed the aspect sentence filtering and observed a drop in the results by 0.59 Rouge-1 F1 score. This demonstrates the effectiveness of aspect-based cluster filtering in improving the overall performance of the model.

4.5 Human Evaluation

We conducted a human evaluation to assess the effectiveness of our model by providing a set of 150 randomly selected papers along with their ground-truth reviews and generated summaries to three domain experts in NLP with a minimum of 5 years of experience. Table 6 shows the results of the evaluation. We asked the responders to evaluate the summaries by rating them between 1 to 10 on Likert Scale (Taherdoost, 2019) based on the following :

- Q1 (Aspect-coverage): Assesses which summary effectively captures the opinions about the specified aspects.
- Q2 (Readability): Evaluates the readability of the summaries.
- Q3 (Diversity): Identifies which summary contains the least amount of repetitive information.
- Q4 (Informativeness): Assesses the usefulness of the summary by providing information about the original reviews.

Consistent with the automated evaluation results, summaries generated by "APCS without aspect" achieved the best scores for Aspect-Coverage, Informativeness, and Diversity compared to the baselines. However, the model may still generate redundant phrases in summaries, particularly in the pros and cons, resulting in a low diversity score. Additionally, the readability score for APCS (both) was lower than that of KL-Sum. The reason for this is that KL-Sum is extractive, meaning that the summaries are taken directly from human-written reviews, while APCS generates abstractive summaries. The readability of BART and APCS (both) is similar. In contrast, the abstractive summary generated by APCS (aspect) effectively captures ideas on aspects. The APCS aspect-based model achieved high Aspect-coverage as it focuses mainly on each aspect of the reviews. However, APCS (both) performed better than PEGASUS on every score, despite both being abstractive methods of summary generation. These results validate the quality of our generation method. We also observed that our model fails when argument is misclassified by the pre-trained model or the aspect classification model makes wrong prdecitions.

5 Conclusion and Future Work

We have proposed a novel architecture for generating both generic and aspect-based pros and cons summaries of peer reviews, utilizing both supervised and unsupervised methods. Our results demonstrate the effectiveness of our proposed architecture. As a future work, investigating the scalability of our proposed architecture for larger datasets and its performance on a diverse range of research domains would also be valuable.

ACKNOWLEDGEMENT

Sandeep Kumar acknowledges the Prime Minister Research Fellowship (PMRF) program of the Govt of India for its support.

References

Nadeem Akhtar, Nashez Zubair, Abhishek Kumar, and Tameem Ahmad. 2017. Aspect based sentiment oriented summarization of hotel reviews. *Procedia Computer Science*, 115:563–571. 7th International Conference on Advances in Computing Communications, ICACC-2017, 22-24 August 2017, Cochin, India.

Phyllis B. Baxendale. 1958. Machine-made index for technical literature - an experiment. *IBM J. Res. Dev.*, 2(4):354–361.

Chaitanya Bhatia, Tribikram Pradhan, and Sukomal Pal. 2020. Metagen: An academic meta-review generation system. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1653–1656, New York, NY, USA. Association for Computing Machinery.

Lutz Bornmann and Hans-Dieter Daniel. 2010. Reliability of reviewers' ratings when using public peer review: a case study. *Learn. Publ.*, 23(2):124–131.

Lutz Bornmann, Markus Wolf, and Hans-Dieter Daniel. 2012. Closed versus open reviewing of journal manuscripts: how far do comments differ in language use? *Scientometrics*, 91(3):843–856.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Ed Collins, Isabelle Augenstein, and Sebastian Riedel. 2017. A supervised approach to extractive summarisation of scientific papers. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada, August 3-4, 2017*, pages 195–205. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13042–13054.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.*, 22:457–479.

Günes Erkan and Dragomir R. Radev. 2011. Lexrank: Graph-based lexical centrality as salience in text summarization. *CoRR*, abs/1109.2128.

Michael Fromm, Evgeniy Faerman, Max Berrendorf, Siddhartha Bhargava, Ruoxia Qi, Yao Zhang, Lukas Dennert, Sophia Selle, Yang Mao, and Thomas Seidl. 2020. Argument mining driven analysis of peer-reviews. In *AAAI Conference on Artificial Intelligence*.

Yanping Fu and Yun Liu. 2022. Domain adaptation with a shrinkable discrepancy strategy for cross-domain sentiment classification. *Neurocomputing*, 494:56–66.

Mahak Gambhir and Vishal Gupta. 2017. Recent automatic text summarization techniques: a survey. *Artif. Intell. Rev.*, 47(1):1–66.

Shima Gerani, Giuseppe Carenini, and Raymond T. Ng. 2019. Modeling content and structure for abstractive review summarization. *Comput. Speech Lang.*, 53:302–331.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1602–1613. ACL.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370, Boulder, Colorado. Association for Computational Linguistics.

Hiroaki Hayashi, Prashant Budania, Peng Wang, Chris Ackerson, Raj Neervannan, and Graham Neubig. 2021. WikiAsp: A Dataset for Multi-domain Aspect-based Summarization. *Transactions of the Association for Computational Linguistics*, 9:211–225.

Xinyu Hua and Lu Wang. 2017. A pilot study of domain adaptation effect for neural abstractive summarization. In *Proceedings of the Workshop on New Frontiers in Summarization, NFiS@EMNLP 2017, Copenhagen, Denmark, September 7, 2017*, pages 100–106. Association for Computational Linguistics.

Hal Daumé III. 2009. Frustratingly easy domain adaptation. *CoRR*, abs/0907.1815.

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. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1647–1661. Association for Computational Linguistics.

Hayato Kobayashi, Masaki Noguchi, and Taichi Yatsuka. 2015. Summarization based on embedding distributions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1984–1989. The Association for Computational Linguistics.

Michail Kovanis, Raphaël Porcher, Philippe Ravaud, and Ludovic Trinquart. 2016. The global burden of journal peer review in the biomedical literature: Strong imbalance in the collective enterprise. *PLoS ONE*, 11.

Sandeep Kumar, Tirthankar Ghosal, Prabhat Kumar Bharti, and Asif Ekbal. 2021. Sharing is caring! joint multitask learning helps aspect-category extraction and sentiment detection in scientific peer reviews. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 270–273.

Florian Kunneman, Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2018. Aspect-based summarization of pros and cons in unstructured product reviews. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2219–2229, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Julian Kupiec, Jan O. Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *SIGIR'95, Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, Washington, USA, July 9-13, 1995 (Special Issue of the SIGIR Forum)*, pages 68–73. ACM Press.

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

HyunSoo Lee, YunSeok Choi, and Jee-Hyong Lee. 2020. Attention history-based attention for abstractive text summarization. In *SAC '20: The 35th ACM/SIGAPP Symposium on Applied Computing, online event, [Brno, Czech Republic], March 30 - April 3, 2020*, pages 1075–1081. ACM.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020b. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Fangtao Li, Chao Han, Minlie Huang, Xiaoyan Zhu, Yingju Xia, Shu Zhang, and Hao Yu. 2010. Structure-aware review mining and summarization. In *COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China*, pages 653–661. Tsinghua University Press.

H. P. Luhn. 1958a. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

Hans Peter Luhn. 1958b. The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2(2):159–165.

Rebecca Mason, Benjamin Gaska, Benjamin Van Durme, Pallavi Choudhury, Ted Hart, Bill Dolan, Kristina Toutanova, and Margaret Mitchell. 2016. Microsummarization of online reviews: An experimental study. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 3015–3021. AAAI Press.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Ani Nenkova, Lucy Vanderwende, and Kathleen R. McKeown. 2006. A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. In *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, August 6-11, 2006*, pages 573–580. ACM.

Makbule Gulcin Ozsoy, Ferda Nur Alpaslan, and Ilyas Cicekli. 2011. Text summarization using latent semantic analysis. *J. Inf. Sci.*, 37(4):405–417.

Daraksha Parveen, Mohsen Mesgar, and Michael Strube. 2016. Generating coherent summaries of scientific articles using coherence patterns. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 772–783. The Association for Computational Linguistics.

Barbara Plank and Reinard van Dalen. 2019. Citetracked: A longitudinal dataset of peer reviews and citations. In *Proceedings of the 4th Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2019) co-located with the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019), Paris, France, July 25, 2019*, volume 2414 of *CEUR Workshop Proceedings*, pages 116–122. CEUR-WS.org.

Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. 2022. Mred: A meta-review dataset for structure-controllable text generation. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2521–2535. Association for Computational Linguistics.

Christian Stab, Tristan Miller, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources using attention-based neural networks. *ArXiv*, abs/1802.05758.

Hamed Taherdoost. 2019. What is the best response scale for survey and questionnaire design; review of different lengths of rating scale / attitude scale / likert scale.

Andrew Tomkins, Min Zhang, and William D. Heavlin. 2017. Reviewer bias in single- versus double-blind peer review. *Proc. Natl. Acad. Sci. USA*, 114(48):12708–12713.

Jelte M. Wicherts. 2016. Peer review quality and transparency of the peer-review process in open access and subscription journals. *PLoS ONE*, 11.

Adhika Pramita Widyassari, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy, and De Rosal Ignatius Moses Setiadi. 2022. Review of automatic text summarization techniques methods. *Journal of King Saud University - Computer and Information Sciences*, 34(4):1029–1046.

Weiran Xu, Chenliang Li, Minghao Lee, and Chi Zhang. 2020. Multi-task learning for abstractive text summarization with key information guide network. *EURASIP J. Adv. Signal Process.*, 2020(1):16.

Min Yang, Qiang Qu, Ying Shen, Qiao Liu, Wei Zhao, and Jia Zhu. 2018. Aspect and sentiment aware abstractive review summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1110–1120, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yinfei Yang, Forrest Sheng Bao, and Ani Nenkova. 2017. Detecting (un)important content for single-document news summarization. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 707–712. Association for Computational Linguistics.

Kaichun Yao, Libo Zhang, Dawei Du, Tiejian Luo, Lili Tao, and Yanjun Wu. 2020. Dual encoding for abstractive text summarization. *IEEE Trans. Cybern.*, 50(3):985–996.

Tiezheng Yu, Zihan Liu, and Pascale Fung. 2021. Adaptsum: Towards low-resource domain adaptation for abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, pages 5892–5904. Association for Computational Linguistics.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2021. Can we automate scientific reviewing? *CoRR*, abs/2102.00176.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2022. Can we automate scientific reviewing? *J. Artif. Intell. Res.*, 75:171–212.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Aspect	**Definition**
Substance	Does the paper contains substantial experiments to demonstrate the effectiveness of proposed methods? Are there detailed result analysis? Does it contain meaningful ablation studies?
Motivation	Does the paper address an important problem? Are other people (practitioners or researchers) likely to use these ideas or build on them?
Clarity	For a reasonably well-prepared reader, is it clear what was done and why? Is the paper well-written and well-structured?
Meaningful Comparison	Are the comparisons to prior work sufficient given the space constraints? Are the comparisons fair?
Originality	Are there new research topic, technique, methodology, or insight?, etc
Soundness	Is the proposed approach sound? Are the claims in the paper convincingly supported?
Replicability	Is it easy to reproduce the results and verify the correctness of the results? Is the supporting dataset and/or software provided?

Table 8: Definition of aspects