Polarity Calibration for Opinion Summarization

Yuanyuan Lei[1], Kaiqiang Song[2], Sangwoo Cho[2], Xiaoyang Wang[2], Ruihong Huang[1], Dong Yu[2]

[1]Texas A&M University [2]Tencent AI Lab, Bellevue, WA

{yuanyuan, huangrh}@tamu.edu

{riversong, swcho, shawnxywang, dyu}@global.tencent.com

Abstract

Opinion summarization is automatically generating summaries from a variety of subjective information, such as product reviews or political opinions. The challenge of opinions summarization lies in presenting divergent or even conflicting opinions. We conduct an analysis of previous summarization models, which reveals their inclination to amplify the polarity bias, emphasizing the majority opinions while ignoring the minority opinions. To address this issue and make the summarizer express both sides of opinions, we introduce the concept of polarity calibration, which aims to align the polarity of output summary with that of input text. Specifically, we develop a reinforcement training approach for polarity calibration. This approach feeds the polarity distance between output summary and input text as reward into the summarizer, and also balance polarity calibration with content preservation and language naturality. We evaluate our Polarity Calibration model (*PoCa*) on two types of opinions summarization tasks: summarizing product reviews and political opinions articles. Automatic and human evaluation demonstrate that our approach can mitigate the polarity mismatch between output summary and input text, as well as maintain the content semantic and language quality[1].

1 Introduction

Opinions are prevalent in various areas, such as social media posts, customer reviews, spoken conversations, argumentative debates, or political matters (Pang et al., 2008; Liu, 2022). Opinion summarization enables automatically generating a brief and informative summary from a large volume of opinions, reviews, or subjective text (Hu and Liu, 2004; Ganesan et al., 2010; Lei and Huang, 2022; Lei and Cao, 2023; Angelidis and Lapata, 2018; Amplayo and Lapata, 2021). The automatic opinion summarization models simplify the extraction of valuable insights from the extensive pool of subjective content, playing a pivotal role in various information access applications, such as digest creation, decision making, product development, or public perception monitoring (Suhara et al., 2020; Amplayo et al., 2021; Iso et al., 2022).

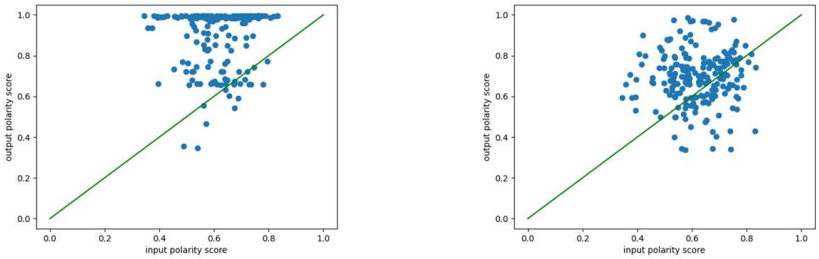
The challenge of opinion summarization lies in presenting divergent or even conflicting opinions. This contrasts sharply with summarizing objective content such as government reports, scientific research, or legal documents, which typically present factual information without the layer of personal perspectives (Erera et al., 2019; Kornilova and Eidelman, 2019; Cachola et al., 2020; Cao and Wang, 2022). Take summarizing product reviews as an example, customers often express differing opinions about the same product, including both positive and negative viewpoints. The central challenge of subjective summarization is aggregating and presenting these disparate opinions.

The critical observation of previously developed summarization models is their tendency to amplify the polarity bias of input text, presenting the majority opinions while ignoring the minority opinions (Section 2). In the example of summarizing product reviews, we quantify the polarity scores of input text and output summaries. Our findings reveal that when the majority of customers express positive opinions about a product, the summarization models directly trained on overwhelming positive reviews can be easily biased to generate overly positive summaries while neglecting the minority of negative opinions (Figure 1, 4). The amplification of polarity bias indicates the limitation in the previous approaches.

To address this issue and proportionally express both sides of opinions, we propose the idea of *polarity calibration*, which aims to align the polarity of output summary with that of input text. In contrast to previous work, we argue that when dealing with conflicting opinions, an intelligent summa-

[1]The code and data link: https://github.com/yuanyuanlei-nlp/polarity_calibration_naacl_2024

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5211–5224



Model generated summaries | Human annotated summaries

Figure 1: The x-axis represents input text polarity score, and the y-axis represents output summary polarity score. The model can amplify the polarity bias, by presenting the majority opinions while ignoring the minority opinions.

rizer should proportionally present both sides of majority and minority opinions, and align with the polarity of input text. Thus, we propose to integrate an additional layer of polarity calibration guidance into the summarizer. The objective of polarity calibration is to encourage the summarizer to exhibit both sides of viewpoints, and mitigate the polarity mismatch between output and input.

To achieve polarity calibration, we develop a reinforcement training approach. More specifically, we employ a polarity reward model to assess the polarity distance between output summary and input text. The polarity distance is incorporated into the summarization model as a reward signal, to encourage the minimization of polarity discrepancy. Besides, to guide the summarizer to maintain the original semantic content of input text, we train a content preservation reward model and feed the content similarity between output and input as reward into the summarizer. In addition, to promote the generation of naturally flowing language, we employ a language naturality reward model and leverage language fluency score as reward. By aggregating the rewards for polarity distance, content preservation, and language naturality, the reinforcement training is designed to balance between improving polarity alignment, retaining content semantic, and generating fluent language.

We evaluate our approach on two types of opinions summarization tasks: summarizing product reviews and political opinions articles. The experiments on both two tasks demonstrate the effectiveness of our method in decreasing the polarity discrepancy between output and input. Both automatic and human evaluation confirm that our approach can enhance polarity alignment, while maintaining content semantic and language quality. Our main contributions are summarized as follows:

- Motivated by the analysis that opinion summarizers tend to amplify the polarity bias, we firstly propose *polarity calibration*, to align the polarity of output summary and input text.
- We design a reinforcement training approach to achieve polarity calibration, by integrating the three rewards for polarity distance, content preservation, language naturality.
- We conduct experiments on two opinions summarization tasks, and effectively decrease the polarity distance while maintaining content semantic and language fluency.

2 Polarity Bias Amplification

This section provides a quantitative analysis of previous summarization models, which reveals their tendency to amplify the polarity bias.

Take product reviews summarization as an example, we aim to examine the polarity of output summary and input reviews. To quantify the polarity, we train a sentiment analysis model on the Amazon product reviews dataset (Keung et al., 2020) to generate polarity scores. This sentiment analyzer is trained to predict whether a review sentence is positive or negative, and we use the predicted probability of the positive class as the polarity score. The polarity score is a numerical value on a scale from zero to one, with zero indicating extreme negative and one indicating extreme positive. Figure 1 illustrate the polarity analysis on the Amazon product reviews summarization dataset (AmaSum) (Bražinskas et al., 2021; Hosking et al., 2023). The x-axis is the average polarity score of sentences in customer reviews, and y-axis is the average polarity score of sentences in summary. Each blue data point represents one product.

The comparison between model generated summaries and human annotated summaries reveals the model's inclination to magnify polarity bias. Figure

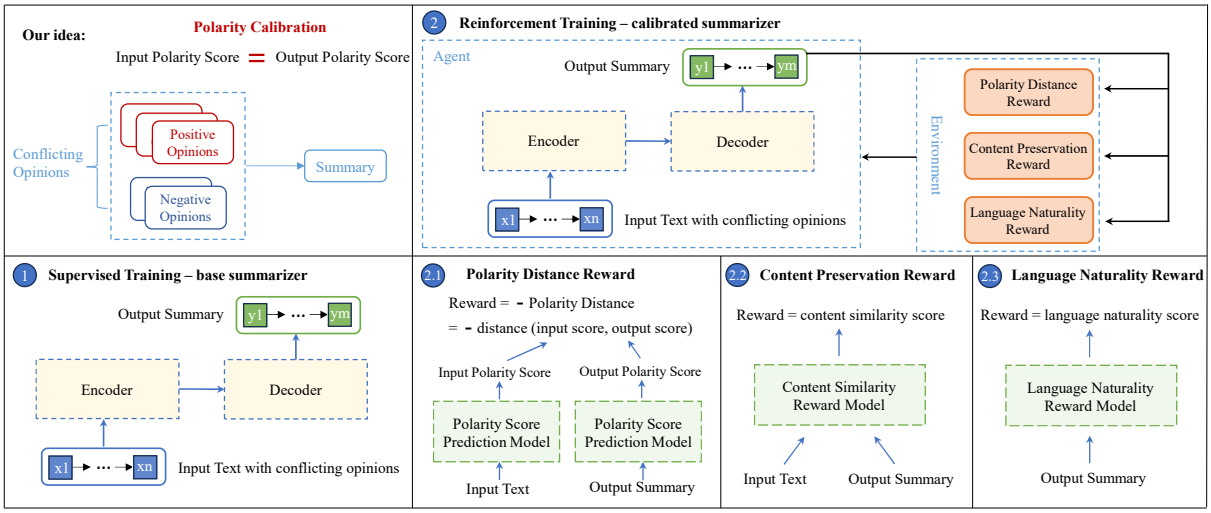


Figure 2: An illustration of polarity calibration with reinforcement learning.

1 takes CopyCat model (Bražinskas et al., 2020) as an example, but this observation also exists in other summarization models (Figure 4). While human consciously maintain the polarity level of input text when crafting summaries, the models tend to generate overly positive summaries, overlooking the minority of negative opinions. One possible explanation is that the models trained on the text predominated with positive reviews tend to develop a bias in favor of highlighting the majority of positive opinions. To guide the model to present both sides of opinions in a proportional manner and better align with the input, we propose to calibrate the polarity score of output summary and input text.

3 Polarity Calibration

This section introduces the methodology for polarity calibration, which is illustrated in Figure 2. The polarity calibration is designed in two steps: firstly training a base summarizer with supervised learning, to equip the model with opinion summarization ability, and secondly training a calibrated summarizer through reinforcement learning, with the aim of refining the model's polarity alignment. The calibrated summarizer after polarity calibration is named as *PoCa*.

3.1 Supervised Training

In the supervised training stage, we train a base summarizer M_{base}, with the ability to summarize opinions. We employ the flan-T5-large model as the backbone model (Chung et al., 2022; Raffel et al., 2023). The input x is the concatenation of different reviews or opinions, denoted as $(x_1, x_2, ..., x_n)$. The goal of the base summarizer is to generate a summary $y = (y_1, y_2, ..., y_m)$, and the human annotated summaries serve as the ground truth labels for training. The cross-entropy loss (CE) is utilized as the learning objective:

$$L_{CE} = \sum_{t=1}^{T} \log \pi_\theta(y_t | y_{t-1}, x) \quad (1)$$

The supervised training makes the base summarizer generate text that is close to the human written reference. However, simply minimizing cross-entropy loss without additional polarity knowledge does not guarantee the polarity alignment between output and input. A generated text that only express the majority opinion can also achieve high Rouge score when compared to human written reference. To imbue the model with polarity awareness, we propose to incorporate an extra guidance of polarity calibration through reinforcement learning.

3.2 Reinforcement Training

In the reinforcement training stage, we train a calibrated summarizer $M_{calibrate}$ on the basis of base summarizer M_{base}. The input x is the concatenation of different reviews or opinions. The goal of the calibrated summarizer is to generate a summary $\hat{y} = (\hat{y}_1, \hat{y}_2, ..., \hat{y}_m)$ that retains the semantics of the input x and calibrates with its polarity level.

We formulate the reinforcement learning for polarity calibration as a system composed of an agent (A), action (a), policy (π), and reward (R). The agent is the summarization model with parameters θ that observes the current state (the model output) at time t and takes an action a (predict the

next word $\hat{y}_t$) by using a policy (π). The reward (R) is a scalar calculated by the reward models $R : \hat{y} \rightarrow [0, 1]$, to evaluate the quality of generated text $\hat{y}$. This reward is then returned as feedback to the summarization model. The objective of reinforcement learning is to maximize the reward (R) by updating the parameters θ of the agent:

$$J(\theta) = E_{\pi_\theta(\hat{y}|x)}[R(\hat{y})] \quad (2)$$

Since the reward is the discrete function of the model's output, the reinforcement learning objective $J(\theta)$ is non-differentiable with respect to the model parameter θ, which makes it difficult to back-propagate the error signals from the reward models to the summarizer. This issue can be addressed through policy gradient (Sutton et al., 1999). Specifically, the expected reward is approximated using a sampling method and the model is trained using stochastic gradient ascent (Williams, 1992), which can be formulated as:

$$\bigtriangledown_\theta J(\theta) = E_{\pi_\theta(\hat{y}|x)}[R(\hat{y}) \bigtriangledown_\theta \log_{\pi_\theta}(\hat{y}|x)] \quad (3)$$

where π_θ is a policy that generates a probability of picking a word as output. The policy gradient learns the optimal policy directly by modifying the model parameters based on the observed rewards.

3.3 Reward Models

In the reinforcement training stage, the generated summary $\hat{y}$ is expected to meet three objectives: (i) reduce polarity distance between output and input (ii) preserve the content semantics of input text (iii) ensure language to be grammatical correct and fluent. Based on the above objectives, the designed reward function $R(x, \hat{y})$ consists of three rewards:

$$R(x, \hat{y}) = \alpha R_P(x, \hat{y}) + \beta R_C(x, \hat{y}) + \gamma R_L(\hat{y}) \quad (4)$$

where $R_P(x, \hat{y})$ is the polarity distance reward calculated between output $\hat{y}$ and input x, $R_C(x, \hat{y})$ is the content semantic similarity reward between output $\hat{y}$ and input x, and $R_L(\hat{y})$ is the language fluency reward for the output text $\hat{y}$. The hyperparameter α, β, γ represent weights for the respective rewards.

3.3.1 Polarity Distance Reward

The purpose of polarity distance reward is to minimize the polarity difference between output summary $\hat{y}$ and input text x. To measure the polarity distance, we build a polarity score prediction model that quantifies the polarity level of a given text. The polarity distance reward $R_P(x, \hat{y})$ is defined as the negative difference between polarity scores of output summary $\hat{y}$ and input text x:

$$R_P(x, \hat{y}) = -|polarity(\hat{y}) - polarity(x)| \quad (5)$$

The polarity score prediction model is tailored to accommodate different tasks. This paper explores two types of opinion summarization tasks: summarizing product reviews with positive or negative opinions, and summarizing political articles with liberal or conservative political stances.

For the Amazon product reviews summarization task, the polarity score prediction model is a sentiment analysis model. A binary classifier is built based on RoBERTa (Liu et al., 2019), to categorize a review sentence into positive or negative. The classifier is trained on the Amazon product reviews dataset (Keung et al., 2020). The polarity score is the predicted probability of the positive class. Since the input text consists of multiple review sentences, the polarity score of input text x is computed as the average of polarity scores assigned to individual review sentence. The polarity score of output summary $\hat{y}$ is computed as the average polarity score of sentences in the summary.

For the political articles summarization task, the polarity score prediction model is a political stance prediction model. A binary classifier is built based on RoBERTa (Liu et al., 2019), to categorize each article into liberal or conservative stance. The classifier is trained on the political stance dataset All-Sides (Baly et al., 2020). The polarity score of each article is the predicted probability of the conservative class. Given the input text comprising multiple articles with different stances, the polarity score of input text x is computed as the average of polarity scores assigned to each individual article. The polarity score of output summary $\hat{y}$ is also the predicted probability of conservative class for the summary.

3.3.2 Content Preservation Reward

The content preservation reward aims to ensure that the information expressed in the input text is retained in the output summary. To quantify the level of content preservation, we build a content similarity reward model to predict the similarity score between output $\hat{y}$ and input x. A RoBERTa based model is used that takes the $(x, \hat{y})$ pair as input and produce a similarity score. This content similarity reward model is trained on the STS-B semantic similarity dataset (Wang et al., 2018). Considering

the raw predicted similarity score ranges from zero to five, we normalize this raw score into the scale of zero to one, and define it as the reward.

$$R_C(x, \hat{y}) = similarity(\hat{y}, x) \quad (6)$$

3.3.3 Language Naturality Reward

The language naturality reward encourages the generated summary $\hat{y}$ to be grammatically correct and natural sounding. To assess the language naturality, we build a language fluency reward model that predicts the language fluency score of the output $\hat{y}$. A binary classifier using RoBERTa is built to predict the generated summary $\hat{y}$ into grammatical correctness or not. This language naturality reward model is trained on the Corpus of Linguistic Acceptability (CoLA) dataset (Warstadt et al., 2018). The language naturality reward is defined as the predicted probability of the grammatical correctness class.

$$R_L(\hat{y}) = fluency(\hat{y}) \quad (7)$$

4 Experiments

4.1 Datasets

We evaluate our approach using two datasets, each focusing on different types of opinions.

AmaSum (Bražinskas et al., 2021) is the Amazon product reviews summarization dataset, which includes product reviews from a wide range of categories. We use the version that contains a maximum of 100 reviews per product for experiments. The dataset collects human written summaries from professional review websites. The annotated summary consists of three portions: verdicts that emphasize the most important points about a product, pros that describe positive details, and cons that states negative aspects. These three portions are concatenated together to form a single summary. We follow the previous work (Hosking et al., 2023) to evaluate the model on the testing set, which contains 50 products from each of the following four common categories: Electronics, Home & Kitchen, Shoes, Sports & Outdoors.

NeuS (Lee et al., 2022) is the political opinions articles summarization dataset, which collects US political news articles from AllSides website. The articles with different political stances that discuss the same event are grouped together as a cluster. Each cluster contains three articles. The dataset also provides an expert written summary for each cluster of articles. We follow the dataset splitting setting released by Lee et al. (2022), which results in 2452 / 307 / 307 news clusters allocated to the train, development, and test sets, respectively.

4.2 Baselines

Summarizing product reviews has attracted research attention for years. The following models are previously developed for summarizing product reviews and implemented as our baselines:

CopyCat (Bražinskas et al., 2020) is an abstractive method by using the hierarchical continuous latent representations to model products and reviews.

BiMeanVAE (Iso et al., 2021) is an abstractive method that encode full reviews as continuous latent vectors, by taking the average or optimizing the combination of review embeddings (COOP).

QT (Angelidis et al., 2021) uses vector quantization to map sentences to a discrete encoding space, and generates extractive summaries by selecting representative sentences from clusters.

SemAE (Basu Roy Chowdhury et al., 2022) is an extractive method that extends the QT method, by relaxing the discretization and encoding sentences as mixtures of learned embeddings.

Hercules (Hosking et al., 2023) develops both extractive and abstractive method, by encoding sentences from customer reviews into a hierarchical latent space and identifying common opinions.

Summarizing political articles with diverse political opinions has a relatively short research history. There are few previously established methods available for comparison. We follow Lee et al. (2022) to compare with the following systems:

LexRank (Erkan and Radev, 2004) is an unsupervised extractive graph-based model that selects sentences based on graph centrality. The nodes are sentences and the edges are weighted with tf-idf.

BART (Lewis et al., 2020) is a multi-document summarization model that fine tunes BART-large on the Multi-News dataset (Fabbri et al., 2019).

Pegasus (Zhang et al., 2020) is an abstractive model that fine tunes Pegasus-large model on the Multi-News dataset (Fabbri et al., 2019).

NeuS (Lee et al., 2022) develops an abstractive summarization method that learns to generate summary in a hierarchical order from title to article.

ChatGPT is a large language model that generates abstractive summaries via prompting. We use the gpt-3.5-turbo version to obtain the summary.

GPT-4 is another large language model that automatically generates abstractive summaries. We use the gpt-4 version to create the summaries. The

	Polarity Distance		Rouge Scores			
	RMSE	MAE	Rouge-1	Rouge-2	Rouge-L	Rouge-Lsum
Human annotated summaries	0.1794	0.1409	-	-	-	-
LexRank (Erkan and Radev, 2004)	0.2772	0.2442	19.91	2.61	12.09	18.27
CopyCat (Bražinskas et al., 2020)	0.3264	0.2907	17.38	1.36	10.95	15.80
BiMeanVAE-avg (Iso et al., 2021)	0.2819	0.2549	21.31	2.00	12.32	19.63
BiMeanVAE-COOP (Iso et al., 2021)	0.2537	0.2189	23.67	2.71	13.96	21.66
QT (Angelidis et al., 2021)	0.2091	0.1609	21.17	1.55	11.36	19.53
SemAE (Basu Roy Chowdhury et al., 2022)	0.2285	0.1786	20.32	1.62	11.35	18.60
Hercules-abstractive (Hosking et al., 2023)	0.2469	0.2167	19.82	2.15	11.71	18.95
Hercules-extractive (Hosking et al., 2023)	0.1888	0.1556	22.89	3.07	12.55	21.44
ChatGPT (gpt-3.5-turbo)	0.2272	0.1875	23.31	2.76	12.99	21.32
GPT-4 (gpt-4)	0.2005	0.1749	23.06	2.60	12.31	21.08
base summarizer (flan-T5-large)	0.2154	0.1782	**29.23**	**5.64**	**17.19**	**26.69**
calibrated summarizer (PoCa)	**0.1824**	**0.1533**	28.44	5.12	16.96	25.92

Table 1: Automatic Evaluation of product reviews summarization on AmaSum dataset. The root mean squared error and mean absolute error between input text polarity score and output summary polarity score are reported.

	Polarity Distance		Rouge Scores			
	RMSE	MAE	Rouge-1	Rouge-2	Rouge-L	Rouge-Lsum
Human annotated summaries	0.1984	0.1517	-	-	-	-
LexRank (Erkan and Radev, 2004)	0.2282	0.1838	38.68	15.94	25.66	33.67
BART (Lewis et al., 2020)	0.2799	0.2291	38.22	15.73	25.52	34.24
Pegasus (Zhang et al., 2020)	0.2810	0.2344	37.33	16.02	25.54	31.45
NeuS (Lee et al., 2022)	0.2172	0.1666	39.09	18.93	29.73	35.35
ChatGPT (gpt-3.5-turbo)	0.2552	0.2076	42.01	16.24	26.12	37.27
GPT-4 (gpt-4)	0.2626	0.2133	42.35	16.48	26.30	37.30
base summarizer (flan-T5-large)	0.2162	0.1613	**43.83**	**20.75**	31.75	**39.16**
calibrated summarizer (PoCa)	**0.1834**	**0.1389**	43.68	20.70	**31.98**	**39.16**

Table 2: Automatic Evaluation of political opinions articles summarization on NeuS dataset. The root mean squared error and mean absolute error between input text polarity score and output summary polarity score are reported.

prompt provided to the model is in Appendix D.

4.3 Automatic Evaluation

The automatic evaluation metrics for polarity calibration are the root mean squared error (RMSE) and mean absolute error (MAE) between polarity scores of output summary and input text. The evaluation metrics for content semantics are calculated by the Rouge scores (Lin, 2004) between model generated summaries and human written reference. The expectation for the summarizer is to minimize polarity distance while maximizing Rouge scores. The results for product reviews summarization on AmaSum dataset are presented in Table 1. The results for political opinions articles summarization on NeuS dataset are shown in Table 2. Our polarity calibration model is named as *PoCa*, and is reported in the last row of tables.

The results demonstrate that polarity calibration through reinforcement training can effectively reduce polarity distance between generated summary and input text while preserving content semantics. When compared to the base summarizer, the calibrated summarizer (PoCa) consistently reduces polarity distance on both AmaSum and NeuS datasets. This indicates that our approach successfully improves polarity alignment between output and input, by incorporating polarity calibration as additional guidance. The statistical t-test indicates a significant difference between calibrated summarizer and base summarizer in terms of polarity distance, but no significant difference in terms of Rouge scores, under the confidence level of 95%. This proves the effectiveness of our approach in mitigating polarity bias without compromising on content semantics.

4.4 Ablation Study

This section studies the effect of three rewards in reinforcement training, and the results on AmaSum dataset is shown in 3. We observe that only feeding the polarity distance reward into the summarizer

	Polarity Distance		Rouge Scores			
	RMSE	MAE	Rouge-1	Rouge-2	Rouge-L	Rouge-Lsum
base summarizer	0.2154	0.1782	**29.23**	**5.64**	**17.19**	**26.69**
+ polarity reward	**0.1545**	**0.1247**	25.24	4.70	15.71	23.05
+ polarity + content reward	0.1839	0.1547	28.13	5.22	16.68	25.78
+ polarity + content + language reward	0.1824	0.1533	28.44	5.12	16.96	25.92

Table 3: The ablation study of three rewards in reinforcement training on AmaSum dataset.

	Polarity	Content		Language	
	Polarity Distance	Non-hallucination	Non-redundancy	Fluency	Coherency
QT (Angelidis et al., 2021)	0.425	0.70	**0.80**	0.30	0.15
Hercules (Hosking et al., 2023)	0.450	0.90	0.25	0.70	0.15
ChatGPT (gpt-3.5-turbo)	0.425	0.90	0.65	0.90	0.90
base summarizer (flan-T5-large)	0.450	0.85	0.75	**0.95**	0.85
calibrated summarizer (PoCa)	**0.350**	**0.95**	**0.80**	**0.95**	**0.95**

Table 4: Human Evaluation of polarity bias, content semantics, and language quality on AmaSum dataset.

can achieve the lowest polarity distance, however, the content semantics is compromised compared to the base summarizer. Learning from both the polarity distance reward and content preservation reward jointly can lead to a reduction in polarity distance while also retaining content semantics. Incorporating the three rewards together can strike a balance between polarity calibration and maintaining content semantics and language quality. This suggests that the three rewards are essential for refining summarization models.

4.5 Human Evaluation

The human evaluation aims to assess the generated summaries from three perspectives: polarity bias, content semantics, and language quality. Specifically, we provide the human annotators with input text and model generated summaries, and ask them five questions regarding polarity bias, content non-hallucination, content non-redundancy, language fluency, and language coherency (Appendix C). Four human annotators who are specialized in natural language processing participated in the evaluation. We select three baselines that achieve low polarity distance, QT (Angelidis et al., 2021), Hercules (Hosking et al., 2023), and ChatGPT, along with our base summarizer and calibrated summarizer for evaluation. To avoid the leakage of model information, different models are randomly shuffled and the name of models are omitted. After collecting the evaluation scores from human annotators for the five questions , we normalize the metrics into the range of zero to one. For the polarity bias evaluation, a lower score is preferable, whereas for the other four questions related to content and language, a higher score is better. The results of human evaluation on AmaSum dataset are presented in Table 4.

The human evaluation showcases the consistent observations with automatic evaluation. The calibrated summarizer exhibits the lowest polarity bias compared to the other models. This proves that the polarity calibration through reinforcement training is effective in mitigating polarity bias. Besides, the calibrated summarizer also achieve high scores for content semantics and language quality. This indicates that the design of content preservation reward and language naturality reward has the effect of improving content and language quality. In summary, our approach enables the mitigation of polarity bias, and also enhances content preservation and language naturality.

4.6 Qualitative Analysis

Figure 3 shows an example of generated summaries from base summarizer and calibrated summarizer. Figure 5 showcases the generated summaries from additional models. We observe that the summary generated by the base summarizer only includes positive opinions without mentioning any negative opinions. This corroborates our finding that the model without polarity calibration tends to develop a bias of emphasizing the majority opinion while overlooking the minority opinion. Nevertheless, the summary produced by the calibrated summarizer not only presents the majority of positive opinions, but also introduces the negative perspective, which aligns better with the input polarity score. This illustrates that polarity calibration integrates polarity awareness into the model, thereby mitigat-

Target polarity score = 0.7225

model	generated summaries	model predicted polarity score	human evaluated **polarity** bias	human evaluation for **content**	human evaluation for **language**
base summarizer	A good choice if you're looking for a waterproof hiking boot that's easy to put on and take off. This boot has a good arch support and a comfortable fit. It's also waterproof, so you won't have to worry about your feet getting wet. A lot of people like the design of this hiking boot, and it's a great choice for those who want a lightweight hiking boot.	0.9825	+ 2	non-hallucination: 1 non-redundancy: 0	fluency: 1 coherency: 1
calibrated summarizer	A good choice if you're looking for a waterproof hiking boot with a wide toe box and good arch support. Waterproof and breathable. Lightweight and comfortable. Good arch support and good traction. Some reports of blisters after a few months of use.	0.7687	0	non-hallucination: 1 non-redundancy: 1	fluency: 1 coherency: 1

Figure 3: The qualitative analysis of generated summaries from base summarizer and calibrated summarizer.

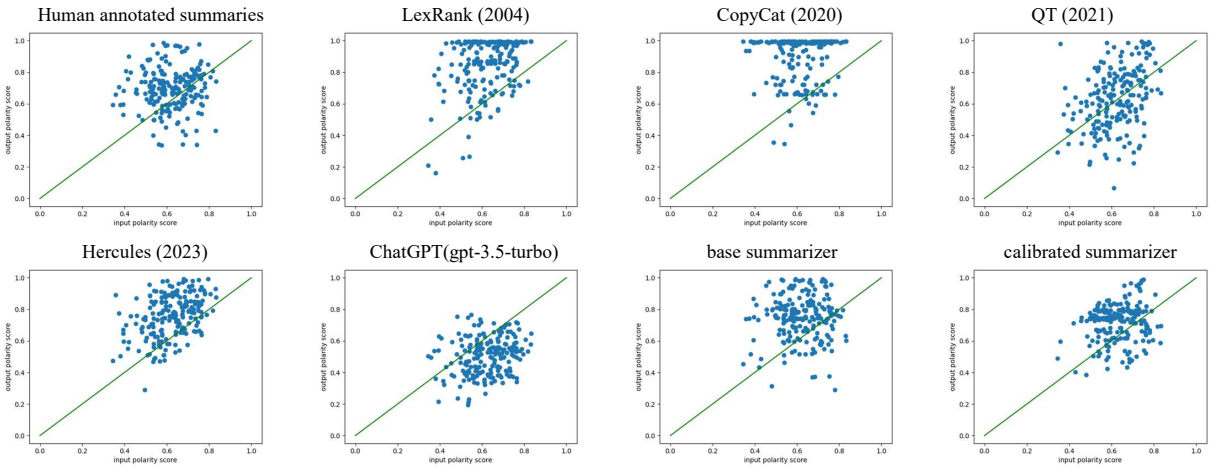


Figure 4: The visualization analysis of generated summaries from various models. The x-axis is input text polarity score, and the y-axis is output summary polarity score. The ideal polarity calibration lies on the green y=x line.

ing polarity bias and enhancing polarity alignment.

4.7 Visualization Analysis

The visualization analysis of generated summaries from different models on AmaSum dataset is shown in Figure 4. The x-axis is input text polarity score, which is calculated as the average polarity score of sentences in input customer reviews. The y-axis is output summary polarity score, which is calculated as the average polarity score of sentences in output summary. Each blue data point represents one product, and the ideal polarity calibration lies on the green y=x line. We observe that while human consciously try to maintain the polarity level of the input text when writing summaries, the models lacking polarity calibration tend to amplify polarity bias, resulting in excessively positive or negative summaries. The calibrated summarizer achieves improved polarity calibration when compared to the base summarizer. This underscores the effectiveness of reinforcement training in refining the model's polarity alignment.

5 Related Work

Opinion Summarization has evolved for years in the natural language processing community. Erkan and Radev (2004) builds a graph to extract the most representative sentences as a summary. Gerani et al. (2014); Di Fabbrizio et al. (2014) rely on text planners and templates. Isonuma et al. (2019) introduces an unsupervised approach for single review summarization. Bražinskas et al. (2020) designs an abstractive method by modeling the hierarchical continuous latent representations. Iso et al. (2021) proposes an optimized combination method to encodes reviews and aggregate review embeddings. Angelidis et al. (2021) maps sentences to a discrete encoding space through vector quantization and extracts the representative sentences from clusters. Basu Roy Chowdhury et al. (2022) develops an extractive method by encoding sentences as mixtures of learned embeddings. Hosking et al. (2023) proposes both extractive and abstractive models by leveraging hierarchical discrete latent space. In contrast to previous work, we aim to address the issue of amplifying polarity bias in opinion summarization models, by incorporating polarity calibration through reinforcement learning.

Bias Mitigation has garnered increasing attention in recent years (Lei and Huang, 2023b,a). The majority of research to address bias mitigation focus on gender bias (Sun et al., 2019) or political bias

(Lei et al., 2022). Manzini et al. (2019) aims to detect and remove multi-class bias in word embeddings. Bordia and Bowman (2019) identifies and reduces gender bias in word-level language models. Recent work devise methods to correct linguistics bias. Pryzant et al. (2020); Madanagopal and Caverlee (2023) reduce linguistic bias by editing text segments such as words or sentences. Liu et al. (2021) introduces a transformer-based model to reduce bias by rewriting biased text. In contrast to these studies, our research investigates the issue of polarity bias in opinion summarization. Our approach aims to mitigate polarity bias and enhance polarity alignment in subjective summarization.

Reinforcement Learning has been frequently used for sequence generation tasks to mitigate exposure bias or to directly optimize task-specific evaluation metrics (Ranzato et al., 2015; Henß et al., 2015; Bahdanau et al., 2016; Paulus et al., 2017; Fedus et al., 2018). In addition, reinforcement learning has been explored for a variety of natural language processing tasks such as question answering (Xiong et al., 2017), knowledge graph reasoning (Lin et al., 2018), relation extraction (Qin et al., 2018), language generation (Li et al., 2016), and text summarization (Chen and Bansal, 2018). Our work develop a reinforcement learning approach to calibrate polarity, by designing rewards for polarity bias, content preservation, and language naturality.

6 Conclusion

This paper focuses on opinion summarization task. We conduct an analysis of previous summarization models, which reveals their tendency to amplify the polarity bias in input text. To mitigate polarity bias and improve polarity alignment between output summary and input text, we introduce the concept of polarity calibration. A reinforcement learning approach is developed for polarity calibration, by designing three rewards for polarity distance, content semantics, and language fluency. Experiments demonstrate the effectiveness of our approach in calibrating polarity while preserving content semantics and language quality.

Limitations

In this paper, we have presented a reinforcement learning-based approach for polarity calibration. To enhance the robustness of our method, future research should investigate the influence of various reward model configurations and alternative reward model designs on polarity calibration. Besides, the experiments focus on summarizing two specific types of opinions, product reviews and political opinions articles. To broaden the scope of our approach and assess its applicability across diverse domains, it would be valuable to examine the effectiveness of polarity calibration in other types of opinion summarization tasks.

Ethical Considerations

This paper investigates the issue of amplifying polarity bias within subjective summarization. The polarity bias is a type of unwanted bias, which hinders the fair representation of both majority and minority opinions in summarization models. The goal of this paper is to mitigate the unwanted polarity bias and enhance polarity alignment in the opinion summarization model. The release of code and model should be leveraged to address and reduce unwanted bias, serving a broader social good.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable feedback and input.

References

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Aspect-controllable opinion summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6578–6593, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Reinald Kim Amplayo and Mirella Lapata. 2021. Informative and controllable opinion summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2662–2672, Online. Association for Computational Linguistics.

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293.

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. *arXiv preprint arXiv:1808.08858*.

Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2016. An actor-critic algorithm for sequence prediction. *arXiv preprint arXiv:1607.07086*.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991, Online. Association for Computational Linguistics.

Somnath Basu Roy Chowdhury, Chao Zhao, and Snigdha Chaturvedi. 2022. Unsupervised extractive opinion summarization using sparse coding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1209–1225, Dublin, Ireland. Association for Computational Linguistics.

Shikha Bordia and Samuel R Bowman. 2019. Identifying and reducing gender bias in word-level language models. *arXiv preprint arXiv:1904.03035*.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169, Online. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2021. Learning opinion summarizers by selecting informative reviews. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9424–9442, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online. Association for Computational Linguistics.

Shuyang Cao and Lu Wang. 2022. HIBRIDS: Attention with hierarchical biases for structure-aware long document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 786–807, Dublin, Ireland. Association for Computational Linguistics.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. *arXiv preprint arXiv:1805.11080*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Giuseppe Di Fabbrizio, Amanda Stent, and Robert Gaizauskas. 2014. A hybrid approach to multi-document summarization of opinions in reviews. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 54–63.

Shai Erera, Michal Shmueli-Scheuer, Guy Feigenblat, Ora Peled Nakash, Odellia Boni, Haggai Roitman, Doron Cohen, Bar Weiner, Yosi Mass, Or Rivlin, Guy Lev, Achiya Jerbi, Jonathan Herzig, Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, Francesca Bonin, and David Konopnicki. 2019. A summarization system for scientific documents. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 211–216, Hong Kong, China. Association for Computational Linguistics.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

William Fedus, Ian Goodfellow, and Andrew M Dai. 2018. Maskgan: better text generation via filling in the_. *arXiv preprint arXiv:1801.07736*.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd international conference on computational linguistics (Coling 2010)*, pages 340–348.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond 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)*, pages 1602–1613.

Stefan Henß, Margot Mieskes, and Iryna Gurevych. 2015. A reinforcement learning approach for adaptive single-and multi-document summarization. In *GSCL*, pages 3–12.

Tom Hosking, Hao Tang, and Mirella Lapata. 2023. Attributable and scalable opinion summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8488–8505, Toronto, Canada. Association for Computational Linguistics.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Hayate Iso, Xiaolan Wang, Stefanos Angelidis, and Yoshihiko Suhara. 2022. Comparative opinion summarization via collaborative decoding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3307–3324, Dublin, Ireland. Association for Computational Linguistics.

Hayate Iso, Xiaolan Wang, Yoshihiko Suhara, Stefanos Angelidis, and Wang-Chiew Tan. 2021. Convex Aggregation for Opinion Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3885–3903, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Masaru Isonuma, Junichiro Mori, and Ichiro Sakata. 2019. Unsupervised neural single-document summarization of reviews via learning latent discourse structure and its ranking. *arXiv preprint arXiv:1906.05691*.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Anastassia Kornilova and Vladimir Eidelman. 2019. BillSum: A corpus for automatic summarization of US legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56, Hong Kong, China. Association for Computational Linguistics.

Nayeon Lee, Yejin Bang, Tiezheng Yu, Andrea Madotto, and Pascale Fung. 2022. NeuS: Neutral multi-news summarization for mitigating framing bias. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3131–3148, Seattle, United States. Association for Computational Linguistics.

Yuanyuan Lei and Houwei Cao. 2023. Audio-visual emotion recognition with preference learning based on intended and multi-modal perceived labels. *IEEE Transactions on Affective Computing*, 14(4):2954–2969.

Yuanyuan Lei and Ruihong Huang. 2022. Few-shot (dis)agreement identification in online discussions with regularized and augmented meta-learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5581–5593, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuanyuan Lei and Ruihong Huang. 2023a. Discourse structures guided fine-grained propaganda identification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 331–342, Singapore. Association for Computational Linguistics.

Yuanyuan Lei and Ruihong Huang. 2023b. Identifying conspiracy theories news based on event relation graph. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9811–9822, Singapore. Association for Computational Linguistics.

Yuanyuan Lei, Ruihong Huang, Lu Wang, and Nick Beauchamp. 2022. Sentence-level media bias analysis informed by discourse structures. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10040–10050, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. 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.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xi Victoria Lin, Richard Socher, and Caiming Xiong. 2018. Multi-hop knowledge graph reasoning with reward shaping. *arXiv preprint arXiv:1808.10568*.

Bing Liu. 2022. *Sentiment analysis and opinion mining*. Springer Nature.

Ruibo Liu, Chenyan Jia, and Soroush Vosoughi. 2021. A transformer-based framework for neutralizing and reversing the political polarity of news articles. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–26.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Karthic Madanagopal and James Caverlee. 2023. Reinforced sequence training based subjective bias correction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2577–2590.

Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047*.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pages 480–489.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018. Robust distant supervision relation extraction via deep reinforcement learning. *arXiv preprint arXiv:1805.09927*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

Yoshihiko Suhara, Xiaolan Wang, Stefanos Angelidis, and Wang-Chiew Tan. 2020. OpinionDigest: A simple framework for opinion summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5789–5798, Online. Association for Computational Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256.

Caiming Xiong, Victor Zhong, and Richard Socher. 2017. Dcn+: Mixed objective and deep residual coattention for question answering. *arXiv preprint arXiv:1711.00106*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

A Implementation

The polarity calibration is implemented within two steps: firstly training a base summarizer with supervised learning to equip the model with opinion summarization ability, and secondly training a calibrated summarizer through reinforcement learning to refine the model's polarity alignment.

In the supervised learning stage, the number of training epochs is set to 10. We use the AdamW (Loshchilov and Hutter, 2019) as the optimizer. The weight decay is set to 1e-2. The batch size is 32. The portion of warm up phase is 0.05. The learning rate is initialized as 1e-5 and adaptively adjusted by a linear scheduler.

In the reinforcement learning stage, the weights α, β, γ assigned to the polarity distance reward, content preservation reward, and language naturality reward in equation (4) are 1.0, 0.5, 0.2 respectively. The weight decay is set to 1e-2. The batch size is 32. The learning rate is set to 1e-6.

B Evaluation of Reward Models

The polarity score prediction model for product reviews summarization is a sentiment analysis model. A binary classifier is built to categorize the text into positive or negative class. The classifier is trained on the Amazon product reviews dataset (Keung et al., 2020). The Precision is 0.9052, Recall is 0.9022, and F1 score is 0.9035.

The polarity score prediction model for political articles summarization is a political stance prediction model. A binary classifier is built to categorize each article into liberal or conservative stance. The model is trained on the political stance dataset All-Sides (Baly et al., 2020). The Precision is 0.8829, Recall is 0.8906, and F1 score is 0.8864.

The content preservation reward model is a content similarity model that predicts a similarity score between two text. The STS-B semantic similarity dataset (Wang et al., 2018) annotates the similarity score for each text pair from 1 to 5. The model is trained with the mean squared error loss function to predict these scores. The Pearson correlation evaluated on the eval set is 0.9109.

The language naturality reward model is a language fluency prediction model. A binary classifier is built to predict the text into grammatical correctness or not. The model is trained on the Corpus of Linguistic Acceptability (CoLA) dataset (Wang et al., 2018). The accuracy on the eval set is 0.8504.

C Human Evaluation

The human evaluation aims to assess the generated summaries from three perspectives: polarity bias, content semantics, and language quality. Specifically, we provide the human annotators with input text and model generated summaries, and ask them the following five questions related to polarity bias, content non-hallucination, content non-redundancy, language fluency, and language coherency.

1. Is the **polarity** of the generated text too positive or too negative compared to the input text? Choose 2, 1, 0, -1, -2. Scores explanation: 2 - far more positive than the target, 1 - a little more positive, 0 - very close, -1 - a little more negative, -2 - far more negative

2. Does the **content** of the generated text hallucinate compared to the reviews? Choose 1 or 0. Scores explanation: 1 - not hallucinate, 0 - has hallucinations

3. Is the **content** of the generated text redundant? Choose 1 or 0. Scores explanation: 1 - concise and not redundant, 0 - has redundancy content

4. Is the **language** of the generated text fluent and grammatically correct? Choose 1 or 0. Scores explanation: 1 - fluent and grammatically correct, 0 - not fluent and has grammar errors

5. Is the **language** of the generated text coherent? Choose 1 or 0. Scores explanation: 1 - coherent, 0 - not coherent

D Prompt for ChatGPT and GPT-4

The prompt provided into gpt-3.5-turbo and gpt-4 baselines for product reviews summarization is "Please summarize the following customer reviews text. Text: <reviews>. Summary:"

The prompt provided into gpt-3.5-turbo and gpt-4 baselines for political articles summarization is "Please summarize the following text. Text: <articles>. Summary:"

E Qualitative Analysis

Figure 5 provides an example of generated summaries from different models. The polarity score of input reviews which is also the target polarity score of output summary is 0.7225. The polarity score of summaries generated by each model is provided.

Target polarity score = 0.7225

model	generated summaries	model predicted polarity score	human evaluated **polarity** bias	human evaluation for **content**	human evaluation for **language**
QT (2021)	My old hiking boots finally bit the dust and good riddance! Used the boots right from the box for a 3-day backpacking trip over 21 miles. I ordered 1/2 size up and with good quality hiking socks they fit perfect. They are amazingly light but supportive at the same time; it is like walking on air. Even bought another pair for the misses.	0.8291	+ 2	non-hallucination: 0 non-redundancy: 1	fluency: 0 coherency: 0
Hercules (2023)	Wore them on a 14-mile hike out of the box. Great ankle support. This is my third pair of Salomon hiking boots. Very waterproof. Comfortable right out of the box. Best boots I have ever owned. Best hiking boots!!!!! Excellent boot. Very comfortable, rugged and provided excellent support. No break in period. These are great boots. Great boot. I love these boots. Great Lacing system.	0.9811	+ 2	non-hallucination: 1 non-redundancy: 0	fluency: 0 coherency: 0
ChatGPT (gpt-3.5-turbo)	The customer mentioned that the boots are lightweight, offer good arch and ankle support, and have a durable construction. They liked the roomy toe box, waterproof feature, and the lacing system of the boots. The customer found them to be versatile and reliable for various activities. They also stated that the boots have a wide toe box and provide excellent grip. The customer is disappointed with the quality of the boots as the rubber part has come unglued after just three weeks of use. The customer also mentions that the boots are not available in wider sizes and that the laces are slippery and refuse to stay tied. The upper of the boots is stiff and uncomfortable, causing blisters. The customer also experienced issues with the tread wearing out quickly and the upper stitching wearing out.	0.5598	- 1	non-hallucination: 1 non-redundancy: 0	fluency: 1 coherency: 1
base summarizer	A good choice if you're looking for a waterproof hiking boot that's easy to put on and take off. This boot has a good arch support and a comfortable fit. It's also waterproof, so you won't have to worry about your feet getting wet. A lot of people like the design of this hiking boot, and it's a great choice for those who want a lightweight hiking boot.	0.9825	+ 2	non-hallucination: 1 non-redundancy: 0	fluency: 1 coherency: 1
calibrated summarizer	A good choice if you're looking for a waterproof hiking boot with a wide toe box and good arch support. Waterproof and breathable. Lightweight and comfortable. Good arch support and good traction. Some reports of blisters after a few months of use.	0.7687	0	non-hallucination: 1 non-redundancy: 1	fluency: 1 coherency: 1

Figure 5: The qualitative analysis of generated summaries from various models.

The summary generated by the calibrated summarizer has the closest polarity score with the input text. The human evaluation results for each model are also provided.