Learning Opinion Summarizers by Selecting Informative Reviews

Arthur Bražinskas[1], Mirella Lapata[1] and **Ivan Titov[1,2]**
[1] ILCC, University of Edinburgh
[2] ILLC, University of Amsterdam
abrazinskas@ed.ac.uk, {mlap,ititov}@inf.ed.ac.uk

Abstract

Opinion summarization has been traditionally approached with unsupervised, weakly-supervised and few-shot learning techniques. In this work, we collect a large dataset of summaries paired with user reviews for over 31,000 products, enabling supervised training. However, the number of reviews per product is large (320 on average), making summarization – and especially training a summarizer – impractical. Moreover, the content of many reviews is not reflected in the human-written summaries, and, thus, the summarizer trained on random review subsets hallucinates. In order to deal with both of these challenges, we formulate the task as jointly learning to select informative subsets of reviews and summarizing the opinions expressed in these subsets. The choice of the review subset is treated as a latent variable, predicted by a small and simple selector. The subset is then fed into a more powerful summarizer. For joint training, we use amortized variational inference and policy gradient methods. Our experiments demonstrate the importance of selecting informative reviews resulting in improved quality of summaries and reduced hallucinations.

1 Introduction

Summarization of user opinions expressed in online resources, such as blogs, reviews, social media, or internet forums, has drawn much attention due to its potential for various information access applications, such as creating digests, search, and report generation (Hu and Liu, 2004; Medhat et al., 2014; Angelidis and Lapata, 2018; Amplayo and Lapata, 2021a). Although significant progress has been observed in supervised summarization in non-subjective context, such as news articles (Rush et al., 2015; Nallapati et al., 2016; See et al., 2017; Lebanoff et al., 2018; Gehrmann et al., 2018; Fabbri et al., 2019; Laban et al., 2020), modern deep learning methods rely on large amounts of annotated data that are not readily available in the

Verdict	If you like the idea of a glass feeder, this is the one to get. It has a lot to offer for the price.
Pros	• Has a large opening that makes it easy to get in and out of the feeder • Has a nice design that's easy to clean
Cons	• The lid is a little flimsy, and it's not as durable as some of the other models
Reviews	... looks just as nice as the glass feeders ‖... Very happy with the value, quality and function ... ‖... the cheapest most flexible "jar" I've ever seen ... ‖... Nice large opening so it's easy to pour the sugar water ‖... This feeder has a nice large opening ... ‖... this is the perfect design and size ... ‖The hummingbirds liked it and had no trouble feeding or perching.... ‖... The main compartment is easy to clean... ‖... The top is a little flimsy ... ‖... it fell out of the hanger it broke for good ... there are so many other nice ones out there that have glass "jar's" or at least sturdier plastic ... ‖... The tray is easy to clean ...

Table 1: Example summary generated by SELSUM with colored alignment to the input reviews. The reviews are truncated and delimited with '‖'.

opinion-summarization domain and expensive to produce. Specifically, the annotated datasets range from 50 to 200 annotated products (Chu and Liu, 2019; Bražinskas et al., 2020b; Angelidis and Lapata, 2018; Angelidis et al., 2020).

The absence of large high-quality resources for supervised learning has called for creative solutions in the past. There is a long history of applying unsupervised and weakly-supervised methods to opinion summarization (Mei et al., 2007; Titov and McDonald, 2008; Angelidis and Lapata, 2018; Chu and Liu, 2019; Amplayo and Lapata, 2020; Bražinskas et al., 2020b).

In this work, we introduce the largest multi-document opinion summarization dataset AMA-

SUM consisting of verdicts, pros and cons for more than 31,000 summarized Amazon products, as shown in Table 1. The summaries were written by professional product reviewers guiding online audience to make better purchasing decisions. In turn, each product is linked to more than 320 reviews, on average. This, however, makes it virtually impossible to train a conventional encoder-decoder model using standard hardware. Moreover, not all reviews cover the summary content. Thus, training to predict summaries based on random review subsets results in hallucinations, as we will empirically demonstrate in Sec. 5.2. This calls for specialized methods selecting smaller subsets of relevant reviews that are fed to a summarizer. We explore this direction by introducing SELSUM that jointly learns to **se**lect and **sum**marize review subsets using *amortized variational inference* and *policy gradient optimization* (Kingma and Welling, 2013; Mnih and Gregor, 2014; Deng et al., 2018), as depicted in Fig. 1.

To select relevant review subsets in training, we utilize the summary to pre-compute lexical features. Then we score review relevance with a tiny neural selector that has only 0.1% of the deep encoder's parameters. These simple features, as opposed to deep encoder representations, allow us to select reviews from large collections without a significant computational burden. Subsequently, only selected reviews are encoded by an 'expensive' encoder, in order to predict the summary. To select quality review subsets in test time, when the summary is not available, we approximate the summary relevance using another neural selector. In our experiments, we show the importance of accurate review selection, affecting the summarizer in training and its output in testing. Furthermore, we show that our model outperforms alternatives in terms of ROUGE scores and content fidelity. All in all, our contributions can be summarized as follows[1]:

- We provide the largest dataset for multi-document opinion summarization;

- We propose an end-to-end model selecting and summarizing reviews;

- We empirically demonstrate superiority of our model to alternatives.

---

[1]The codebase and dataset are available at https://github.com/abrazinskas/SelSum.

2 Dataset

The dataset (AMASUM) is based on summaries for consumer products written by professional reviewers, in English. We focused on four main professional product review platforms: bestreviews.com (BR); cnet.com; pmag.co.uk (PM); runrepeat.com (RR). The former three mostly offer content for electronic consumer products, while the last one for sport shoes. These summaries provide a quick-glance overview of a product to help users make informed purchases. Unlike customer reviewers on public platforms, such as Amazon, professional reviewers concentrate on quality writing and deliberately utilize many sources of information. These sources include reading customer reviews on public platforms, making online research, asking expert users for an opinion, and testing products themselves. In general, the summaries come in two forms. The first ones are verdicts, usually a few sentences emphasizing the most important points about a product. The second ones are pros and cons, where the most important positive and negative details about a product are presented. These tend to be more detailed, and focus on fine-grained product aspects, such as Bluetooth connectivity, resolution, and CPU clock speed.

As content providers compete for online users, the summaries **are** what the user wants as opposed to what researchers **believe** the user wants. This is in contrast to crowd-sourcing where researchers bias the worker writing process with assumptions about what constitutes a good summary. The assumptions are rarely verified by a marketing research or user testing. In turn, this has lead to a large variance of summary styles and composition even in the same domain (Angelidis and Lapata, 2018; Chu and Liu, 2019; Bražinskas et al., 2020b,a).

2.1 Content Extraction

We wrote HTML scraping programs for each platform and extracted segments containing verdicts, and pros and cons. Further, from advertisement links we extracted Amazon standard identification numbers (ASINs) which allowed us to identify what Amazon products are reviewed and link summaries to the Amazon product catalog.

We used various paid services to obtain Amazon reviews and product metadata. We fetched verified reviews for all products, and utilized unverified
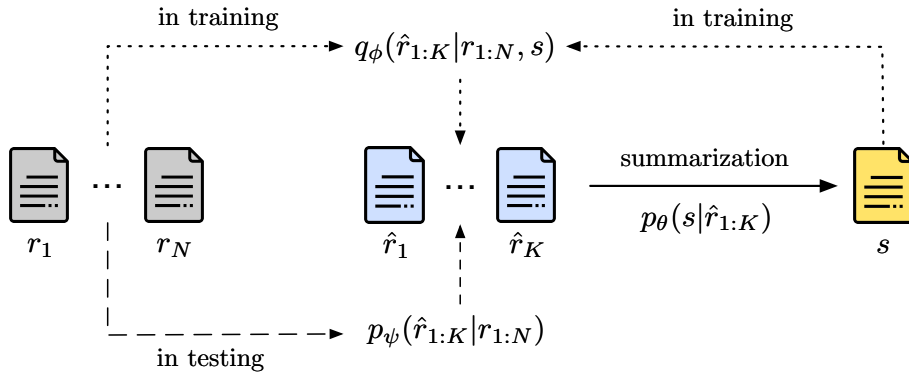
Figure 1: The SELSUM model is trained to select and summarize a subset of relevant reviews $\hat{r}_{1:K}$ from a full set $r_{1:N}$ using the approximate posterior $q_\phi(\hat{r}_{1:K}|r_{1:N}, s)$. To yield review subsets in test time, we fit and use a parametrized prior $p_\psi(\hat{r}_{1:K}|r_{1:N})$.

ones only for unpopular products (< 50 reviews). We also utilized a publicly available Amazon review dataset (Ni et al., 2019).

2.2 Filtering

We removed all reviews that have less than 10 and more than 120 words. We also removed all unpopular products that have less than 10 reviews. Further, we removed all summaries that have less than 5 words, and all instances that have either verdict or pros or cons missing. The overall statistics comparing our final dataset to available alternatives are shown in Table 2. Our dataset is substantially larger than the alternatives, both in terms of number of summaries and their associated reviews.

2.3 Summary Statistics

We analyzed summaries from different platforms in terms of their lengths and ROUGE recall with respect to reviews, as shown in Table 3. First of all, verdicts tend to be shorter than pros and cons, and concentrate on fewer aspects. They also exhibit higher word overlap to user reviews as indicated by higher ROUGE scores. We also observed that pros and cons tend to concentrate on specific product features, which can often be found in product meta information (product description, the bullet list of features). Cons tend to be shorter than pros, we believe, primarily because most summarized products are rated highly (4.32/5.0 on average).

3 Approach

As summaries are written mostly for popular products, with more than 320 reviews on average, it is computationally challenging to encode and attend all the available ones to decode the summary. To alleviate this problem, we propose to condition the decoder on a smaller subset of reviews. However, as not all reviews provide a good content coverage of the summary. Thus, training on random subsets leads to hallucinations, as we will show in Sec. 5.2. Instead, we propose to learn a review selector, which chooses reviews guided by the summary. We frame this as a latent variable modeling problem (the selection is latent) and rely on the variational inference framework to train the selector, see Sec. 3.2. The selector (the approximate posterior) is a neural module assessing the review-summary relevance using pre-computed lexical features, thus, efficiently selecting from large review collections. Further, the selected reviews are decoded to the summary, as illustrated in Fig. 1. To select reviews in test time, we train a review selector that does not rely on the summary, as presented in Sec. 3.3.

3.1 Probabilistic Framing

Let $\{r_{1:N}^i, s^i\}_{i=1}^M$ be reviews-summary pairs, and let $\hat{r}_{1:K}$ be a reduced subset of reviews, where $K < N$, and each variable follows a categorical distribution. As review subsets $\hat{r}_{1:K}$ are unknown in advance, they are latent variables in our model, and both the full set $r_{1:N}$ and the summary s are observed variables. To maximize the log-likelihood shown in Eq. 1, we have to marginalize over all possible review subsets.

$$\log p_\theta(s|r_{1:N}) = \log \mathop{\mathbb{E}}_{\hat{r}_{1:K} \sim p(\hat{r}_{1:K}|r_{1:N})} [p_\theta(s|\hat{r}_{1:K})]$$
(1)

Unfortunately, the marginalization is intractable, and thus we leverage the Jensen's inequality (Boyd and Vandenberghe, 2004) to obtain the lower bound as shown in Eq. 2, which, in turn, is approximated

	Ent	Rev/Ent	Summaries (R)	Type	Domain
AMASUM (This work)	31,483	326	33,324 (1.06)	Abs.	Products
SPACE (Angelidis et al., 2020)	50	100	1,050 (3)	Abs.	Hotels
COPYCAT (Bražinskas et al., 2020b)	60	8	180 (3)	Abs.	Products
FEWSUM (Bražinskas et al., 2020a)	60	8	180 (3)	Abs.	Businesses
MEANSUM (Chu and Liu, 2019)	200	8	200 (1)	Abs.	Businesses
OPOSUM (Angelidis and Lapata, 2018)	60	10	180 (3)	Ext.	Products

Table 2: Statistics comparing our dataset to alternatives; R stands for the number of references. For our dataset, we show the average number of reviews and references per entity. We count verdicts, pros and cons of a product as one summary.

	Verdict			Pros			Cons		
	Len	R1	R2	Len	R1	R2	Len	R1	R2
BR (27,329)	20.60	82.40	34.45	37.34	79.12	29.75	16.27	82.19	33.58
CNET (2,717)	29.74	81.05	34.72	32.08	77.85	30.04	25.11	75.16	25.84
PM (1,756)	30.23	76.08	28.28	20.78	65.53	16.09	14.33	62.08	13.81
RR (1,522)	77.86	60.45	13.12	120.04	59.44	13.47	43.36	63.11	16.02
All (33,324)	24.47	80.95	33.18	39.82	77.40	28.31	18.12	79.69	31.10

Table 3: Summary statistics of the dataset. The number of data points is in parentheses.

via Monte Carlo (MC).

$$\log \mathbb{E}_{\hat{r}_{1:K} \sim p(\hat{r}_{1:K}|r_{1:N})} [p_\theta(s|\hat{r}_{1:K})] \geq \\ \mathbb{E}_{\hat{r}_{1:K} \sim p(\hat{r}_{1:K}|r_{1:N})} [\log p_\theta(s|\hat{r}_{1:K})] \quad (2)$$

Here the latent subset $\hat{r}_{1:K}$ is sampled from a prior categorical distribution agnostic of the summary. From the theoretical perspective, it can lead to a large gap between the log-likelihood and the lower bound, contributing to poor performance (Deng et al., 2018). From the practical perspective, it can result in the input reviews not covering the summary content, thus forcing the decoder in training to predict 'novel' content. Consequently, this leads to hallucinations (Maynez et al., 2020) in test time, as we empirically demonstrate in Sec. 5.2.

3.2 Model

To address the previously mentioned problems, we leverage *amortized inference* reducing the gap (Kingma and Welling, 2013; Cremer et al., 2018). And re-formulate the lower bound using weighted sampling as shown in Eq. 3.

$$\log \mathbb{E}_{\hat{r}_{1:K} \sim p(\hat{r}_{1:K}|r_{1:N})} [p_\theta(s|\hat{r}_{1:K})] \geq \\ \mathbb{E}_{\hat{r}_{1:K} \sim q_\phi(\hat{r}_{1:K}|r_{1:N},s)} [\log p_\theta(s|\hat{r}_{1:K})] - \quad (3) \\ \mathbb{D}_{KL} [q_\phi(\hat{r}_{1:K}|r_{1:N},s)||p(\hat{r}_{1:K}|r_{1:N})]$$

The first term, known as *reconstruction*, quantifies the summary prediction quality with re-

view subsets selected by the approximate posterior $q_\phi(\hat{r}_{1:K}|r_{1:N},s)$. Unlike the prior, it selects reviews relevant to the summary s, thus providing a better content coverage of the summary. Hence, it reduces the amount of 'novel' content the decoder needs to predict. As we empirically demonstrate in Sec. 5.2, this results in summaries with substantially fewer hallucinations in test time. The second term, the Kullback-Leibler divergence (KLD), serves as a regularizer preventing the posterior from deviating from the prior. We did not find it useful – presumably because the latent space of our model (i.e. the choice of reviews to summarize) has already very limited capacity – and do not use the KLD term in training. Instead, after training, we fit a rich prior (see Sec. 3.3).

3.2.1 Approximate Posterior

The distribution assigns a probability to every possible subset of reviews $\hat{r}_{1:K}$. However, this would require us to consider $N!/(N-K)!K!$ possible combinations to normalize the distribution (Koller and Friedman, 2009). To make it computationally feasible, we assume a local, left-to-right factorization (Larochelle and Murray, 2011), reducing the complexity to $\mathcal{O}(K\,N)$:

$$q_\phi(\hat{r}_{1:K}|r_{1:N},s) = \prod_{k=1}^{K} q_\phi(\hat{r}_k|r_{1:N},\hat{r}_{1:k-1},s). \quad (4)$$

Technically, each local distribution can be computed by softmax normalizing scores produced by the *inference network* $f_\phi(\hat{r}_k, r_{1:N}, s)$. To represent $(\hat{r}_k, r_{1:N}, s)$ input tuples, we use pre-computed lexical features, such as ROUGE scores for $(\hat{r}_k, s)$ and $(\hat{r}_k, r_{1:N})$, and aspect-coverage metrics (see Appendix 8.7 and Section 5.3). This, in turn, allows us to learn feature inter-dependencies and score large collections of reviews in a fast and memory efficient-manner.

To avoid duplicate reviews, we assume that $\hat{r}_k$ can be any review in the full collection $r_{1:N}$ except previously selected ones in the partial subset $\hat{r}_{1:k-1}$. To accommodate that, we 'block' scores for all previously selected reviews $\hat{r}_{1:k-1}$ as $f_\phi(\hat{r}_k, r_{1:N}, s) = -\inf \forall \hat{r}_k \in \hat{r}_{1:k-1}$. In practice, we compute logits once for $r_{1:N}$, and then perform a progressive distribution re-normalization by 'blocking' logits for previously selected reviews.

3.2.2 Reconstruction

In training, we optimize parameters only for the reconstruction term in Eq. 3. However, this optimization is not straightforward as it requires backpropagation through categorical samples $\hat{r}_{1:K}$ to compute a gradient estimate. Furthermore, it is not possible to apply the re-parametrization trick (Kingma and Welling, 2013) for categorical variables. On the other hand, the Gumbel-Softmax trick (Jang et al., 2017), in its standard form, would require encoding and backpropagating through all possible review subsets, making it computationally infeasible. Instead, we used REINFORCE (Williams, 1992) that considers only a sampled subset for gradient estimation,[2] as shown in Eq. 5. The notation is simplified to avoid clutter.

$$\nabla_\phi \mathop{\mathbb{E}}_{\hat{r}_{1:K} \sim q_\phi(\hat{r}_k | r_{1:N}, \hat{r}_{1:k-1}, s)} [\log p_\theta(s | \hat{r}_{1:K})] =$$
$$\mathop{\mathbb{E}}_{\hat{r}_{1:K} \sim q_\phi} [(\log p_\theta(s | \hat{r}_{1:K}) - \beta(s)) \nabla_\phi \log q_\phi]$$
(5)

Here $\beta(s)$ corresponds to a baseline reducing the gradient variance (Greensmith et al., 2004). Specifically, we used an MC estimate of Eq. 2 by randomly sampling review subsets. Moreover, we were separately updating the posterior and summarizer, in the spirit of stochastic inference (Hoffman et al., 2013). In turn, this made it computationally possible to further reduce the variance by estimating Eq. 5 with more samples.

---

[2] We provide further discussion contrasting REINFORCE and the Gumbel-Softmax trick in Appendix 8.1.

3.3 Fitting a Prior

The selector used in training (i.e the approximate posterior $q_\phi(\hat{r}_{1:K} | r_{1:N}, s)$) cannot be used in test time, as it has a look-ahead to the summary s. Instead we need a prior $p_\psi(\hat{r}_{1:K} | r_{1:N})$. Since we have not used any prior in training (i.e. ignored the KLD term, Eq. 3), we, similarly in spirit to Razavi et al. (2019), fit a parameterized prior after training the summarizer, and then use the prior as the test-time review selector.

Intuitively, the fitted prior tries to mimic predictions of the approximate posterior without having access to s. We care only about the mode of the distribution, so, to simplify the task, we select the most likely review subset from the posterior to train the test time selector and frame it as a binary prediction task. Let $\{r_{1:N}^i, s^i\}_{i=1}^M$ be reviews-summary pairs where we utilize $q_\phi(\hat{r}_{1:K} | r_{1:N}, s)$ to create $\{r_{1:N}^i, d_{1:N}^i\}_{i=1}^M$ pairs. Here, d_j is a binary tag indicating whether the review r_j was selected by the posterior. This dataset is then used to train the score function $f_\psi(r_k; r_{1:N})$. In test time, we select K reviews with the highest scores.

We score reviews with a binary classifier that inputs review semantic representations. The representations are computed in two steps. First, we independently encode reviews word by word, then compute the weighted average of word representations. Second, we pass all $r_{1:N}$ averaged representations through another encoder (contextualizer) to capture review interdependence features. Details can be found in Appendix 8.3.

4 Experimental Setup

4.1 Data Preprocessing

In our experiments, we used a preprocessed version of the dataset described in Sec. 2. First, we set the full review set size N to 100 maximum reviews, and the review subset size K was set to 10 entries. Further, we split the dataset to 26660, 3302, and 3362 summaries for training, validation, and testing, respectively. For our models training, verdicts, pros, and cons were joined to one sequence with a separator symbol indicating boundaries.

4.2 Baselines

LEXRANK (Erkan and Radev, 2004) is an unsupervised extractive graph-based model that selects sentences based on graph centrality. Sentences represent nodes in a graph whose edges are weighted with tf-idf.

MEANSUM (Chu and Liu, 2019) is an unsupervised abstractive summarization model which treats a summary as a structured latent state of an autoencoder trained to reconstruct reviews of a product.

COPYCAT (Bražinskas et al., 2020b) is the state-of-the-art unsupervised abstractive summarizer with hierarchical continuous latent representations to model products and individual reviews.

RANDOM: here we split all N reviews by sentences, and randomly selected 3, 7, 4 sentences for verdicts, pros, and cons, respectively.

EXTSUM: we created an extractive summarizer trained on our data. First, we used the same ROUGE greedy heuristic as in Liu and Lapata (2019) to sequentially select summarizing verdict, pro, and con sentences from the full set of reviews using the actual gold summary (ORACLE). Further, we trained a model, with the same architecture as the prior in Sec. 3.3, to predict sentence classes. More details can be found in Appendix 8.2.

4.3 Alternative Review Selectors

To better understand the role of review selection, we trained the same encoder-decoder summarizer as in SELSUM but with two alternative selectors.

Random reviews We trained and tested on random review subsets (RANDSEL). Here, review subsets were re-sampled at each training epoch.

ROUGE-1 top-k We produced review subsets based on review-summary ROUGE-1 R scores (R1 TOP-K) for training.[3] Specifically, we computed the scores for each pair, and then selected K reviews with highest scores to form the subset. To select reviews in test time, we trained a selector as in Sec. 3.3.

4.4 Experimental details

Below we briefly describe model details; more information can be found in Appendix 8.5.

Summarizer We used the Transformer encoder-decoder architecture (Vaswani et al., 2017) initialized with base BART (Lewis et al., 2020b), 140M parameters in total. Reviews were independently encoded and concatenated states of product reviews were attended by the decoder to predict the summary as in Bražinskas et al. (2020a). We used

ROUGE-L as the stopping criterion. Summary generation was performed via the beam search of size 5 and with 3-gram blocking (Paulus et al., 2017).

Posterior For the inference network in Sec. 3.2.1, we used a simple non-linear two-layer feed-forward network with 250 hidden dimensions. The model consisted of 95k parameters. The network inputs 23 pre-computed features. For instance, ROUGE-1 and -2 scores between each review and the summary, and each review and other reviews in the full set. Similar to Ni et al. (2019), we tagged fine-grained aspect words to compute precision and recall scores between reviews and the summary, and used them as features.

Prior For the parametrized prior in Sec. 3.3, we used fine-tuned encoders on the end-task from both R1 TOP-K and SELSUM. For the contextualizer we used a cold-start Transformer encoder with 2 layers and 8-head attention mechanisms. For score networks, we used 2 hidden layer feed-forward networks with the ReLU non-linearities and 100 hidden dimensions. Dropouts at each layer were set to 0.10. In total, the model had 97M parameters. The details of the architecture can be found in Appendix 8.3.

Pros and cons classification COPYCAT and MEANSUM are not specifically designed for pros and cons generation. Therefore, we used a separately trained classifier to split each summary to pros and cons.

Extractive summarizer We used a pre-trained BART encoder, and 100 hidden states for 1 layer score feed-forward network with ReLU, and 0.1 dropout. The contextualizer had one layer, and the final score feed-forward had 100 hidden dimensions, 0.1 dropout, with layer normalization before logits are computed. We trained the model for 5 epochs, with 1e-05 learning rate.

Automatic evaluation We separately evaluated verdicts, pros, and cons with the standard ROUGE package (Lin, 2004)[4], and report F1 scores.

Human evaluation To assess content support, we randomly sampled 50 products, generated summaries, and hired 3 workers on Amazon Mechanical Turk (AMT) for each HIT. To ensure high qual-

---

[3]We tried but were not able to obtain better results by turning the scores into a distribution and sampling from it, so we used the deterministic strategy in the main experiments.

[4]We used a wrapper over the package https://github.com/pltrdy/files2rouge.

ity submissions, we used qualification tasks and filters. More details can be found in Appendix 8.4.

Hardware All experiments were conducted on 4 x GeForce RTX 2080 Ti.

5 Results

5.1 Automatic Evaluation

The results in Table 4 suggest that the supervised models substantially outperform the unsupervised ones. Also, all supervised abstractive summarizers outperform EXTSUM, suggesting recombining information from reviews into fluent text is beneficial. Among the summarizers with the review selectors, SELSUM yields the best results on verdicts and cons. Although, we noticed that SELSUM generates shorter pros than R1 TOP-K, which may harm its scores (Fan et al., 2018)[5]. Further, when random reviews were used both in training and testing (RANDSEL), the results are substantially lower. On the other hand, when review subsets were produced by SELSUM and summarized by RANDSEL (marked with '*'), we observed a substantial increase in all the scores. This suggests the importance of deliberate review selection in test time. In general, all models yield higher scores on pros than cons, which is expected as most reviews are positive (on average 4.32/5) and it is harder for the model to find negative points in input reviews.

5.2 Content Support

Generating input faithful summaries is crucial for practical applications, however, it remains an open problem in summarization (Maynez et al., 2020; Fabbri et al., 2020; Wang et al., 2020). Moreover, ROUGE scores were shown not always be reliable for the content support assessment (Tay et al., 2019; Bražinskas et al., 2020b). Therefore, we evaluated generated summary sentences via AMT, as in Bražinskas et al. (2020b), using the following options.

Full support: all the content is reflected in the reviews; *Partial support*: only some content is reflected in the reviews; *No support*: content is not reflected in the reviews.

First, we observed that random reviews in training and testing (RANDSEL) lead to summaries with a significant amount of hallucinations. Further, when RANDSEL summarizes reviews chosen by

SELSUM's selector ('the prior') – indicated by '*' – the content support is still substantially lower than with SELSUM. This demonstrates that having a selection component is necessary not only at test time but also in training; without it, the model does not learn to be faithful to input reviews. Lastly, SELSUM generates substantially more input faithful summaries than R1 TOP-K.

5.3 Posterior-Selected Review Subsets

We performed extra experiments to understand why SELSUM model performs better than R1 TOP-K. Recall, their difference is only in the review selector used in training. SELSUM learns a neural model as the posterior, whereas R1 TOP-K relies on a ROUGE-1 heuristic. We hypothesize that SELSUM exploits more expressive features (beyond ROUGE-1) to select reviews that are more relevant to the summary, helping SELSUM to learn a stronger model, less prone to hallucinations.

In order to validate this, in Table 6 we show their results on the test set but in the training regime, i.e. with reviews selected while accessing the actual gold summary. As in training, R1 TOP-K uses the ROUGE-1 heuristic, while SELSUM relies on the learned posterior. Naturally, both methods obtain stronger scores in this artificial set-up (Table 6 vs. Table 4). What is more interesting is that SELSUM is considerably stronger than R1 TOP-K, suggesting that the SELSUM's selection component indeed chooses more relevant reviews.

Lastly, to rank each feature 'importance', we estimated the mutual information (MI) (Kraskov et al., 2004; Ross, 2014) between the posterior input features and the binary decision to select a review, as in Sec. 3.3. We found that besides review-vs-summary ROUGE-1 and -2 scores, the posterior uses fine-grained aspect features, and review-vs-all-reviews ROUGE scores (quantifying the uniqueness of each review). See also Appendix 8.7.

6 Related Work

Due to a lack of annotated data, extractive weakly-supervised opinion summarization has been the dominant paradigm. LEXRANK (Erkan and Radev, 2004) is an unsupervised extractive model. OPINOSIS (Ganesan et al., 2010) does not use any supervision and relies on POS tags and redundancies to generate short opinions. Although, it can recombine fragments of input text, it cannot generate novel words and phrases and thus produce

---

[5] R1 TOP-K and SELSUM generate 31.95 and 27.14 words on average, respectively.

	Verdict			Pros			Cons		
	R1	R2	RL	R1	R2	RL	R1	R2	RL
ORACLE	38.14	11.76	31.50	37.22	10.53	33.50	34.09	10.75	29.66
RANDOM	13.12	0.82	10.85	14.29	1.04	13.02	9.91	0.72	8.77
LEXRANK	15.12	1.84	12.60	14.12	1.50	12.81	8.28	0.82	7.24
MEANSUM	13.78	0.93	11.70	10.44	0.63	9.55	5.95	0.45	5.29
COPYCAT	17.05	1.78	14.50	15.12	1.48	13.85	6.81	0.82	5.89
EXTSUM	18.74	3.01	15.74	19.06	2.47	17.49	11.63	1.19	10.44
RANDSEL	23.25	4.75	17.82	20.26	3.60	18.52	13.59	2.32	11.86
RANDSEL*	23.95	5.16	18.49	21.06	3.94	19.31	13.78	2.35	12.10
R1 TOP-K	23.43	4.94	18.52	**22.01**	3.94	**19.84**	14.93	2.57	12.96
SELSUM	**24.33**	**5.29**	**18.84**	21.29	**4.00**	19.39	**14.96**	**2.60**	**13.07**

Table 4: Test set ROUGE F1 scores on verdict, pros and cons. The last block shows review selection variants, where RANDSEL* was trained on random review subsets but tested on SELSUM-selected subsets.

	Verdict			Pros			Cons		
	Full↑	Partial↓	No↓	Full↑	Partial↓	No↓	Full↑	Partial↓	No↓
RANDSEL	28.96	45.90	25.14	38.62	29.10	32.28	14.92	14.60	70.48
RANDSEL*	50.79	31.75	17.46	50.62	22.96	26.42	16.84	**13.75**	69.42
R1 TOP-K	55.21	31.77	13.02	56.07	26.61	17.31	33.33	27.78	38.89
SELSUM	**66.08**	**25.15**	**8.77**	**70.21**	**17.99**	**11.80**	**38.41**	29.21	**32.38**

Table 5: Human evaluated content support. Percentages are based on summary sentences. RANDSEL* was trained on random review subsets but tested on SELSUM selected subsets.

	Verdict	Pros	Cons
	RL	RL	RL
R1 TOP-K	19.38	**21.09**	13.26
SELSUM	**20.44**	20.79	**14.40**

Table 6: Test set ROUGE F1 scores when review selection is guided by the gold summary.

coherent abstractive summaries. Other earlier approaches (Gerani et al., 2014; Di Fabbrizio et al., 2014) relied on text planners and templates, which restrict the output text. A more recent method of Angelidis and Lapata (2018) applies multiple specialized models to produce extractive summaries. More recently, there has been a spark of interest in unsupervised abstractive opinion summarization. Such models include MEANSUM (Chu and Liu, 2019), COPYCAT (Bražinskas et al., 2020b), DE-NOISESUM (Amplayo and Lapata, 2020), OPINIONDIGEST (Suhara et al., 2020), and CONDA-SUM (Amplayo and Lapata, 2021b).

Our work is related to the extractive-abstractive summarization model (Chen and Bansal, 2018) that selects salient sentences from an input document using reinforcement learning. They assume one-to-one mapping between extracted and summary sentences for news. In opinion summarization, however, we often need to fuse user opinions expressed in multiple reviews. Lastly, unlike their model, our selector and summarizer are trained jointly to predict the summary using a differentiable loss. Also, our model is related to the unsupervised paraphrasing MARGE model (Lewis et al., 2020a), where the decoder has a modified attention mechanism accounting for the target-source document similarity. However, in their approach, the actual selection of relevant documents is performed offline via heuristics. This, in turn, makes it non-differentiable and over-reliant on the modified attention mechanism. We, however, learn the selector (posterior) jointly with summarizer, and select reviews in the online regime.

An alternative to review subsets selection are more memory and computationally efficient attention mechanisms (Beltagy et al., 2020; Pasunuru et al., 2021). However, it is unclear what relationship exists between attention weights and model outputs (Jain and Wallace, 2019), thus, making it harder to offer evidence for generated summaries. In our case, the summarizer relies only on a selected subset and generates summaries faithful to

its content.

In general, in news summarization, which is a more mature branch, large datasets are commonly obtained from online resources (Sandhaus, 2008; Hermann et al., 2015; Grusky et al., 2018; Narayan et al., 2018; Fabbri et al., 2019). The most relevant dataset is MULTINEWS Fabbri et al. (2019), where journalist-written summaries are linked to multiple news articles. The most similar opinion summarization dataset SPACE (Angelidis et al., 2020) contains 1050 summaries produced for 50 hotels by crowd-sourcing.

7 Conclusions

In this work, we introduce the largest multi-document abstractive dataset for opinion summarization. The dataset consists of verdicts, pros and cons, written by professional writers for more than 31,000 Amazon products. Each product is linked to more than 320 customer reviews, on average. As standard encoding-decoding is computationally challenging, we perform summarization with an integrated component that selects smaller review subsets. We conclude that the 'naive' selection of random reviews leads to content infidelity (aka hallucinations) and present SELSUM that learns to select and summarize reviews end-to-end. The model is computationally efficient, scaling to large collections. Its summaries result in better ROUGE scores and are better supported by input reviews.

7.1 Ethics Statement

Human Evaluation We used a publicly available service (Amazon Mechanical Turk) to hire voluntary participants, requesting native speakers of English. The participants were compensated above minimum hourly wage both in the USA and the UK, the self-reported locations of the participants.

Dataset The dataset was collected and used in accordance to the non-commercial purpose. The dataset is intended for non-commercial and educational purposes only. It will be made available free of charge for these purposes without claiming any rights, similar to (Grusky et al., 2018). To maintain privacy, all summary writers are anonymized.

Acknowledgments

References

Reinald Kim Amplayo and Mirella Lapata. 2020. Unsupervised opinion summarization with noising and denoising. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1934–1945, Online. Association for Computational Linguistics.

Reinald Kim Amplayo and Mirella Lapata. 2021a. 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.

Reinald Kim Amplayo and Mirella Lapata. 2021b. 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. 2020. Extractive opinion summarization in quantized transformer spaces. In *In Transactions of the Association for Computational Linguistics (TACL)*.

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686, Brussels, Belgium. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Stephen Boyd and Lieven Vandenberghe. 2004. *Convex optimization*. Cambridge university press.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020a. Few-shot learning for opinion summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4119–4135, Online. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020b. 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.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of ACL*.

Eric Chu and Peter Liu. 2019. Meansum: a neural model for unsupervised multi-document abstractive summarization. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 1223–1232.

Chris Cremer, Xuechen Li, and David Duvenaud. 2018. Inference suboptimality in variational autoencoders. In *International Conference on Machine Learning*, pages 1078–1086. PMLR.

Yuntian Deng, Yoon Kim, Justin Chiu, Demi Guo, and Alexander M. Rush. 2018. Latent alignment and variational attention. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9735–9747, Red Hook, NY, USA. Curran Associates Inc.

Giuseppe Di Fabbrizio, Amanda Stent, and Robert Gaizauskas. 2014. A hybrid approach to multi-document summarization of opinions in reviews. pages 54–63.

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.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2020. Summeval: Re-evaluating summarization evaluation. *arXiv preprint arXiv:2007.12626*.

Angela Fan, David Grangier, and Michael Auli. 2018. Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54, Melbourne, Australia. Association for Computational Linguistics.

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, Beijing, China. Coling 2010 Organizing Committee.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

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)*, pages 1602–1613, Doha, Qatar. Association for Computational Linguistics.

Evan Greensmith, Peter L Bartlett, and Jonathan Baxter. 2004. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5(9).

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. 2013. Stochastic variational inference. *Journal of Machine Learning Research*, 14(5).

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. ACM.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Daphne Koller and Nir Friedman. 2009. *Probabilistic graphical models: principles and techniques*. MIT press.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. 2004. Estimating mutual information. *Physical review E*, 69(6):066138.

Philippe Laban, Andrew Hsi, John Canny, and Marti A. Hearst. 2020. The summary loop: Learning to write abstractive summaries without examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5135–5150, Online. Association for Computational Linguistics.

Hugo Larochelle and Iain Murray. 2011. The neural autoregressive distribution estimator. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 29–37. JMLR Workshop and Conference Proceedings.

Logan Lebanoff, Kaiqiang Song, and Fei Liu. 2018. Adapting the neural encoder-decoder framework from single to multi-document summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4131–4141, Brussels, Belgium. Association for Computational Linguistics.

Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. 2020a. Pre-training via paraphrasing. In *Proceedings of advances in neural information processing systems*.

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.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020. Dice loss for data-imbalanced NLP tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476, Online. Association for Computational Linguistics.

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.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web*, pages 171–180. ACM.

Andriy Mnih and Karol Gregor. 2014. Neural variational inference and learning in belief networks. In *International Conference on Machine Learning*, pages 1791–1799. PMLR.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar GuÌ‡lçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. 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)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.

Ramakanth Pasunuru, Mengwen Liu, Mohit Bansal, Sujith Ravi, and Markus Dreyer. 2021. Efficiently summarizing text and graph encodings of multi-document clusters. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4768–4779, Online. Association for Computational Linguistics.

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

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163, Valencia, Spain. Association for Computational Linguistics.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. 2019. Generating diverse high-fidelity images with vq-vae-2. In *Advances in neural information processing systems*.

Brian C Ross. 2014. Mutual information between discrete and continuous data sets. *PloS one*, 9(2):e87357.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

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.

Wenyi Tay, Aditya Joshi, Xiuzhen Zhang, Sarvnaz Karimi, and Stephen Wan. 2019. Red-faced ROUGE: Examining the suitability of ROUGE for opinion summary evaluation. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 52–60, Sydney, Australia. Australasian Language Technology Association.

Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web*, pages 111–120. ACM.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 83–92.

8 Appendices

8.1 REINFORCE vs Gumbel-Softmax

In our experiments, we used REINFORCE (Williams, 1992) instead of a straight-trough Gumbel-Softmax estimator (Jang et al., 2017), which is a popular alternative. In our case, we need to sample without replacement each $\hat{r}_k$ from the collection $r_{1:N}$. The Gumbel-Softmax, requires to relax the system for 'soft' samples used in the backward pass. In this way, the system is updated considering all possible assignments to the categorical variable, even though only one was sampled in the forward pass. For instance, one could encode all reviews $r_{1:N}$ and weigh their word contextualized representations to obtain each $\hat{r}_k$. However, this is a computationally expensive and memory demanding operation. On the other hand, REINFORCE does not require this relaxation, and the encoder is exposed only to one possible assignment to each $\hat{r}_k$, both in the forward and backward pass.

8.2 Extractive Summarizer

As mentioned in Sec. 4.2, our extractive summarizer had the same architecture as the prior in Sec. 3.3. We independently encoded sentences from reviews, contextualized them, and computed their distributions for 4 classes.

In training, we considered up to 550 sentences, where only up to 16 have positive labels (4, 8, 4 for verdicts, pros, cons, respectively) marked by ORACLE. However, this resulted in label imbalance, where, in training, the model is incentives to ignore positive labels (Li et al., 2020). However, in test time, we care about positive instances only. To counter this problem, we scaled each positive class loss by 50. In this way, the model is forced to prioritize the positive classes more.

At test time, we sequentially selected top-k summarizing sentences for verdicts, pros, and cons. To make each sentence selected either for verdict, pros, or cons, we were sequentially excluding selected sentences from the pool of candidates.

8.3 Prior Score Function

Below we describe the architecture of the score function used in Sec. 3.3, and it is also illustrated in Fig. 2. First, we initialized with a fine-tuned review encoder from the summarizer that was trained using a review selector (i.e. SELSUM or R1 TOP-K). The encoder produces contextualized word representations for each review independently. The word rep-
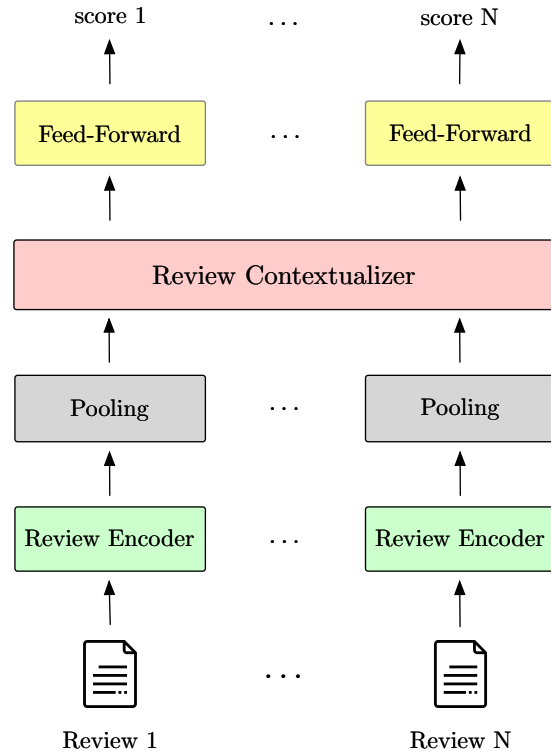

Figure 2: Architecture of the prior score function.

resentations are obtained from the last Transformer layer. Then, we computed the weighted average of these representations to get the review representation. Further, we passed the review representations through another encoder that contextualizes them by attending representations of other reviews in the collection. Finally, we projected the outputs to scores.

8.4 Human Evaluation Setup

To perform the human evaluation experiments described in Sec.5.2, we hired workers with 98% approval rate, 1000+ HITS, from the USA and UK, and the maximum score on a qualification test that we had designed. We payed them 17.25 $ per hour, on average. The task was a minimal version of the actual HIT, where we could test that workers correctly understood the instructions. Also, we asked them if they were native English speakers.

8.5 Experimental Details

Summarizer We used the Transformer encoder-decoder architecture (Vaswani et al., 2017) initialized with base BART (Lewis et al., 2020b), 140M parameters in total. Reviews were independently encoded and concatenated states of product reviews were attended by the decoder to predict the sum-

mary as in Bražinskas et al. (2020a). We used trainable length embeddings, and BPE (Sennrich et al., 2016) vocabulary of 51,200 subwords. Subword embeddings were shared across the encoder and decoder for regularization (Press and Wolf, 2017). For summary generation, we used beam search with the size of 5 and 3-gram blocking (Paulus et al., 2017). Parameter optimization was performed using Adam (Kingma and Ba, 2014) with 5,000 warm-up steps. We trained SELSUM, R1 TOP-K, and RANDSEL for 8, 8, and 9 epochs, respectively. All with the learning rate of 3e-05.

Posterior For the inference network in Sec. 3.2.1, we used a simple two-layer feed-forward, 250 hidden dimensions, with the tanh non-linearity and layer normalization before a linear transformation to scores. The model consisted of 95k parameters. We used 23 static features by treating verdicts and pros and cons as separate summaries. For instance, ROUGE-1 and -2 scores between each review and the summary, and each review and other reviews in the full set. Similar to Ni et al. (2019), we tagged fine-grained aspect words to compute precision and recall scores between reviews and the summary, and used them as features. Full details about features can be found in Appendix 8.7. Lastly, we used 3 samples for the expectation estimation in Eq. 5 and 3 samples to compute the baseline.

Prior For the parametrized prior in Sec. 3.3, we used fine-tuned encoders on the end-task from both R1 TOP-K and SELSUM. For the contextualizer we used a cold-start Transformer encoder with 2 layers, and 8-head attention mechanisms. For score networks, we used 2 hidden layer feed-forward networks with the ReLU non-linearities, and dropouts set to 0.10. In total, the model had 97M parameters. We trained the one for SELSUM and R1 TOP-K for 5 and 4 epochs, respectively, with the 1e-05 learning rate and 5,000 warmup steps. The details of the architecture can be found in Appendix 8.3.

8.6 Aspect-based Metric

In addition to standard unweighted word-overlap metrics commonly used to analyze datasets (Grusky et al., 2018; Fabbri et al., 2019), we also leveraged an aspect specific metric. Similar to Ni et al. (2019), we applied a parser (Zhang et al., 2014) to the training set to yield *(aspect, opinion, polarity)* tuples. From the tuples, we created a lexicon, which contains fine-grained aspect keywords, such as battery life, screen,

Feature	MI
R2-R($\hat{r}$, pc)	0.0634
R1-R($\hat{r}$, pc)	0.0564
R2-P($\hat{r}$, pc)	0.0523
R1-R($\hat{r}$, v)	0.0489
R2-R($\hat{r}$, v)	0.0449
R2-P($\hat{r}_k$, r_{-k})	0.0411
R2-P($\hat{r}$, v)	0.0405
AR($\hat{r}$, pc)	0.0353
R1-R($\hat{r}_k$, r_{-k})	0.0346
AP($\hat{r}$, pc)	0.0331
R2-R($\hat{r}_k$, r_{-k})	0.0313
R1-P($\hat{r}$, pc)	0.0266
R1-P($\hat{r}$, v)	0.0208
AP($\hat{r}_k$, r_{-k})	0.0190
AR($\hat{r}_k$, r_{-k})	0.0173
LD($\hat{r}$, pc)	0.0167
AP($\hat{r}$, v)	0.0151
LD($\hat{r}$, v)	0.0146
R1-P($\hat{r}_k$, r_{-k})	0.0138
AR($\hat{r}$, v)	0.0135
AD($\hat{r}$)	0.0106
AD(v)	0.0005
AD(pc)	0.0003

Table 7: Full list of features sorted by their mutual information to a review being selected to the subset binary variable.

resolution, etc. In addition, to reduce noise, we manually cleaned the lexicon from aspect unrelated keywords, resulting in 2,810 entries. Further, we used the lexicon to automatically tag aspect keywords in text. Lastly, we computed *aspect precision* (AP) and *aspect recall* (AR) scores by comparing two sequences. These scores were used as features in SELSUM.

8.7 Posterior Features

In total, we used 23 continuous features computed for each tuple $(s, \hat{r}_k, \hat{r}_{1:N})$. We can group features into three categories. The first ones were computed for a sequence standalone. The second ones were computed as $f(\hat{r}_k, s)$ where $\hat{r}_k$ is the current review (hypothesis) and s is the summary (reference). The last ones were computed with respect to other reviews as $f(\hat{r}_k, r_{-k})$, where r_{-k} (reference) are all reviews except $\hat{r}_k$ (hypothesis). We also treated verdicts (v) and pros and cons (pc) as separate sequences.

Aspect precision, recall, and density are calculated by leveraging the lexicon presented in Appendix 8.6. Additionally, aspect density (AD) was

computed as the number of unigram aspect keywords, divided by the number of unigrams in a sequence. Finally, length difference LD$(\cdot, \cdot)$ was computed as the difference of two normalized lengths. The Normalization was performed by the maximum sequence length division.

To gain a deeper insight into the SELSUM posterior's inner-workings, we analyzed feature importance for including a review in the subset. Same as in Sec. 3.3, we used the trained posterior to create a binary tagging dataset. Further, we estimated the mutual information (MI) (Kraskov et al., 2004; Ross, 2014) between the posterior input features and the binary decision variable. This allowed us to identify the dependency strength between each feature and the variable.

As features are computed separately for verdicts and pros and cons, we observed that features for pros and cons (pc) have higher MI than for verdicts (v), which suggests that review selection is guided by pros and cons more than by verdicts. Second, fine-grained aspect keyword based scores (AP and AR) also have high MI for pc. This is unsurprising, as pros and cons are often more detailed, making them less predictable based on the prefix, thus the model favours reviews with matching aspect keywords. Lastly, the ROUGE scores computed against other reviews in the collection r_{-k} have high MI. This indicates reliance on global statistics computed based on the whole set of reviews.

8.8 Error Analysis

Human written pros and cons, besides summarizing customer opinion expressed in reviews, can contain details that can often be found in the product meta data. For example, users rarely mention that the same product comes in different colors. Consequently, the decoder is trained to predict such phrases based on the prefix instead of the input reviews, as shown in Example 9. We further observed that cons are harder, in general, to align to reviews.

Human-written summaries sometimes contain customer opinion quantification expressed in phrases, such as "some users" and "a few customers". We observed this to be challenging for the summarizer to generate accurately as shown in Example 10. Especially, it applies to cons that summarize opinions of a small number of users. Logically, such reviews are hard to retrieve from a large collection in training, consequently, the model learns

to rely on local statistics (the prefix). Overall, quantification of user opinions adds an additional layer of complexity for the decoder as besides generating summaries that are content supported in terms of opinions, it needs to quantify the them correctly. This is an interesting future direction for abstractive opinion summarization. Lastly, we observed that online users, in their reviews, sometimes compare the product to other products on the market. This, in turn, can confuse the model and make it generate the summary that contains fragments describing another product. Occasionally, we observed such mistakes in the output.

Verdict	A comprehensive study guide for those who are new to the ASWB exam.
Pros	Offers a variety of practice questions to help you get the most out of the exam. Offers an easy-to-understand overview of the test and how it works.
Cons	The practice questions are not as detailed as the actual exam, and some questions may not be relevant to the actual questions.
Review 1	This review guide claims to reflect the 2018 blueprint of the ASWB exam. However, all ethics questions refer to a 2006 version of the NASW Code of Ethics. The code of ethics is a substantial part of the exam and many of the questions and answer explanations do not reflect what will be on the test. It is not worth the money.
Review 2	I passed my LMSW test on the first try with this as the only study material!!! I would definitely recommend the book for its content and practice test. I will say that the actual test is very different from this practice test in the book as the real tests involves more questions what do you do FIRST and what do you do NEXT? This guide makes it pretty easy to narrow down to two answers whereas the actual test is not that easy. Study all the content and know it! Supervisory and ethics played a big part in this test. Be more prepared for simple application questions than objective content...
Review 3	I used this book as my primary study material for the LMSW licensing exam. While it is a thick book with a lot of information, it was helpful in preparing for the exam. There were some questions on the exam that were not in the book, but I still passed the exam with the information I studied from the book. While the exams are not all the same, I can not guarantee the same results for everyone who uses this book, but I do not have any negative reviews. I bought the book used so I did not utilize the app that comes with it.
Review 4	I just graduated with my MSW in May, and studied with this book as well as the pocket prep app for about a month. This book is a great comprehensive overview of material that we have all learned, and was great for reviewing. The practice questions were also helpful in figuring out HOW the exam wants you to answer. I passed the LMSW exam on my first try today! I will definitely buy the clinical version when I take that in a few years!
Review 5	I read through this book and utilized the practice exam at the end. While this book does go over some foundational content which is applicable to the exam, overall the content is redundant and irrelevant. The practice exam in the back of this book is extremely different than the practice exam offered through the ASWB or the actual exam itself. Multiple licensed social workers I have spoken to have stated that the practice exam offered through the ASWB was the most helpful thing in preparing for the actual exam and understanding how its questions are formatted, which I 100% agree with. For the money I spent on this book, it was disappointingly ineffective as an exam prep tool.
Review 6	Yes, I did it. I have an LCSW(not to be confused with the LCSW-clinical license as in MA the 'C' stands for certified). With this comprehensive and thurough study guide i was able to pass my exam the first time. I felt so prepared after using this. I would say, pair this study guide with an online prep app, as the exam is computer based and using a phone app you will get into the patterning needed to succeed in the exam.
Review 7	This study guide for the ASWB exam is great. It has a lot of review material and a practice test. There is also a code to put on a phone/tablet. I used this and passed on the first time. I also used the BSW guide for that exam and passed on the first time. A must have for anyone looking to pass the Master's Exam.
Review 8	This was a super purchase! It offers excellent tips and strategies to prepping for this challenging exam. It conditioned me to understand the method of the questions and not just knowledge. I just passed the first time! This was incredible because I trained in the UK and not USA. This study kit prepared me to pass what better review can one give?
Review 9	I passed!! This was a great study guide for me. I was intending to read the whole thing but it was a lot so I went through the table of contents and highlighted the sections I wanted to study. It also helped to read it and write it down for memorization. The practice test was hard but it really tests you on your knowledge so don't take it until you are ready. I used this and a few other practice exams.
Review 10	I cannot attest to results as of yet. But I can say that the book has a very organized layout and presents information about how the test is setup which provides great insight for one's approach to testing.

Table 8: Example summary generated by SELSUM with color highlighted content alignment.

Verdict	If you're looking for a set of glass containers that are both BPA-free and dishwasher safe, this is the one to get.
Pros	Glass containers come in a variety of sizes and colors, so you can find the right size for your needs. The lids are easy to open and close. BPA and phthalate-free.
Cons	The containers are on the smaller side, which can make them difficult to store in the microwave.
Review 1	These containers are fantastic. The lips snap on very securely but are extremely easy to remove. I do wash the lids by hand b/c they are top rack dishwasher safe only but that's not a big deal for me b/c the glass can go through the dishwasher. Plus I am saving dishwashing time anyway b/c I previously stored leftovers in Tupperware that I could not heat in the microwave so would have to transfer to a different dish before heating. Now it's a 1 stop shop. Variety of sizes are great as well.
Review 2	These containers are excellent!! One of the things I love about them is that I can fill them all the way to the top and it won't spill out when I put the lid on. There are a variety of sizes included and I use these every day. I only wish they were etched on the bottom with the volume that each bowl can hold. But in any case they are worth every penny. Glass containers don't discolor in the microwave and you don't have to worry about consuming plastic. The lids add a layer of sturdiness to the bowls and I store them in the cupboard with their lids on.
Review 3	Pro: Glass containers can go into the microwave to reheat leftovers without cooking food oils and colors into plastic. Don't use the lids in the microwave. And DO follow the instructions and wash before using, definitely! Slightly con: The rubber gasket will separate from the lid and stick to the glass container if you snap the lid on while it's slightly wet. Maybe if it's completely dry as well! But it's easy enough to peel off and reseat in the lid.
Review 4	Great quality! Nice, secure fitting lids. It's so easy to know what is being stored in the bowls. I have not used them in the freezer. I have only used them in the microwave and refrigerator. We have used every single bowl at one time or another! It's great to have different sizes to accommodate different portions. I would like to get a couple of even larger sizes if I can find them! This glass won't crack as my plastic containers did (unless I happen to drop!). An interlocking system between the bottom of a container and the lid of another, so they would stack more securely, would be REALLY NICE.
Review 5	These are the perfect size for everything, but I'm sad that the blue rubber isn't staying on the lids at all. We are avoiding washing them in dishwasher so they don't get worse. Kinda bummed. UPDATE 6/3/2018 We received the new and improved 1790 Glass Container Set & the modifications made by the manufacturer were a home run! The lids fit tightly and evenly on the containers that using them is a snap. They are so well made. Completely air tight and leak proof. Very impressed with how quickly the issue was addressed and resolved. 5 Star product! If you're looking for versatile containers that deliver..... look no further. They are right here.
Review 6	These are such great dishes! I don't eat a lot so they are perfect for single serving cooking and storage. Love the way the lids clip in place, making a very good seal to keep the food fresh. Baking and cleaning are so easy, just the way I like things, nice and easy. Would recommend this set to anyone looking for a set of small versatile baking/storage options!
Review 7	Some of the flaps on the lids are a little hard to close, but I am guessing that has more to do with the fact that they are new more than anything else. Overall, this is a good, quality set at a great price. Durable in oven and microwave and washes up easily. No staining, bubbles or potential melting issues like plastic containers.
Review 8	I should have read a little closer and counted the actual dishes in the photo - its a NINE piece set unless you count the lids - which you can't store food in. They might as well call it a 27 piece set because of the nine lid seals, which is realistically another 'piece'. The quality is average along with price when you discover how many actual storage containers are sold. So, read the entire description and count the dishes in the photo..
Review 9	Great glass based meal containers. The caps are plastic, and the air seal is a rubber-like material. Works as intended, however the seal part can be separated from the cap and has a tendency to adhere more onto to the glass over the cap. If this happens, be careful when separating the seal and container, as it can cause the air seal to rip a bit.
Review 10	This product is durable and easy to clean. I love that it's BPA free and oven, microwave, freezer, and dishwasher safe. It's air tight and I haven't had anything leak even when putting liquid in. For the price you get, they're great storage containers with versatility in different temperatures. I got this for my boyfriend and will probably buy more for myself if I needed glass containers in the future.

Table 9: Example summary generated by SELSUM with color highlighted errors. As indicated in red, pros can contain details that one would expect to find in product meta data instead of customer reviews. In orange, we indicate a logical mistake. In cyan we indicate a contradiction.

Verdict	If you're looking for a reliable, water-resistant, and weather-resistant G-Shock watch, this is the one to get.
Pros	Solar powered. Water-resistant. Includes atomic clock and countdown timer. Includes a stopwatch and 5 alarms.
Cons	Some owners say the watch is too small for their needs. A few owners say it doesn't have the features of other models.
Review 1	It will become your everyday watch and you will enjoy it. It does everything in the description and then some. It sits nicely on your wrist without looking too big or too small, just right. It has stopwatch capabilities along with timer capabilities. Its a good digital watch and from what I've seen it can take a beating and still beep on the hour for you. Of course there's a couple accessories you can get for it too, including a screen protector and a brace to protect it. If you're looking for a good watch that you can drag through the mud and still check the time, look no further!
Review 2	This Casio G-Shock adjusted itself to the correct time, day of the week and date as soon as I unboxed it and light struck the solar charger. Very easy to set up, unlike some Casio products that require long sequences of button pushing. It meets my needs: updates the time and date automatically via an atomic clock signal, is solar powered, is water resistant, and displays the important information at a glance. Glad I paid a little more for this model than some of the Casios that have thick operating manuals and lots of button pushing to adjust. Well worth the money, and Amazon did a great job with prompt delivery of the correct product on top condition.
Review 3	OK - so it won't give you weather, or headings, or your email... but come on, it sets itself to the freakin' atomic clock EVERY SINGLE DAY. AND it's shock resistant, and water resistant, and solar powered, and lights itself up when you turn it towards your face, AND it has a stop watch, AND it can tell you the time in multiple time zones. SERIOUSLY, this watch is a sleeper, FANTASTIC value for the money.
Review 4	As a purchaser of Casio' G-Shock watches for almost four decades, I have depended on their toughness in rugged environments. This latest update has all the whistles and bells of the bigger versions, but more compact. The atomic-solar G-Shocks seem to last ten years with accurate time and no problems with the battery. I recommend this watch for those who want a minimum of hassle. (Note: I rarely use backlight illumination.)
Review 5	I like the atomic clock sync and multi-time zone options. The UTC time needs to be accurate and easy to get to for celestial navigation. This watch has all that plus many other features and, of course, it's nearly impossible to break and self-charging. This is the perfect watch for someone that is outdoors or on the ocean for extended periods. I highly recommend this for anyone that needs a completely reliable and accurate time piece.
Review 6	This is one of the best value for the money G-Shocks out there. A homage to the original square G-Shock, this watch combines retro styling with modern technology like automatic time setting. Every feature works flawlessly, and it looks and feels good on the wrist. If you're strapped for cash go for the 5600 series, but if you can swing it, this is the entry level G-Shock that will leave you wanting for nothing. Easy to mod or awesome as is.
Review 7	It's cheap, reliable, durable, and fashionable. It's so light it feels like you're not wearing a wrist watch. It tells the time, date, day of the week. It adjusts itself to the atomic clock nearest to you. It's solar powered. You don't have to adjust it or buy a new battery for it (for around a decade!). It has a stopwatch and 5 alarms although I don't see myself using them. It can quickly adjust to different time zones. You can use it it in the dark. I don't know if I listed all the advantages...
Review 8	I love these watches. It is flippin' solar powered with multi-band atomic correction! You can't beat that for maintenance-free operation. I appreciate the simple design and layout without the cluttered display or bulk of other models. A simple, durable, reliable wrist watch.
Review 9	If your interested in a no flash, timeless, simplistic design then this watch is for you. This G-Shock is a step up from Casio classic DW-6500. It add solar charging, world time, 5 alarms, and the ability to sync with 6 different stations around the world. As a fan of Casio's Square G-Shocks this one in particular is one of my favs. It keeps that classic look and durability while improving on it.
Review 10	This model is considered a must have by most G-Shock aficionados. It's the classic G-Square model, but updated with solar power and atomic clock time sync features. It also has the World Time feature. It's a rock-solid watch that is nearly indestructible.

Table 10: Example summary generated by SELSUM with highlighted errors. In this summary, the system incorrectly generated cons with quantifiers.

Gold

Verdict	If you need to keep rodents away from several small places, this is a good option.
Pros	Available in packs of 1, 2, 3, or 4. Deters rodents of all kinds. Safe to use around kids and pets - no noticeable sound. Comes with a built-in night light. Easy to use - just plug in any room where pests are detected.
Cons	Does not repel insects, only rodents. Takes about a week to work.

Generated

Verdict	If you are looking for an inexpensive way to attract and control rodents, this 3-pack of nightlights is a good choice.
Pros	Three nightlights for indoor or outdoor use. Lightweight, compact, and easy to store. Can be used indoors or outdoors.
Cons	Not as effective against rodents as it is for trapping them. Some reports of the nightlights not working.

Selected Reviews

Review 1	Description and photo showed 2 nightlights. Package arrived with one nightlight and 2 unlit units. Kinda annoying Overall, does what it's designed to do.
Review 2	Decent product. Rarely see anything anymore only thing is it messes with the frequency of your belongings like microwave toaster etc my advice is keep them low level but overall nice.
Review 3	So far, these have worked great for ants actually. We haven't seen any critters at all.
Review 4	I can't say how well they work as far as rodents go, but I have yet to see one EVER in my house. I love the little light it illuminates on the floor. It is fantastic at night and gives me piece of mind for little critters.
Review 5	We have not had any more mice in our garage!
Review 6	These little babies actually seem to work! We live in an area with lots of pests. I put 6 of these around the house and we haven't had any issues whatsoever.
Review 7	Bought one of these devices and used while at winter home. When we returned see no sign of critters where we installed.
Review 8	I can't believe it seems to drive them away! Happy Camper.
Review 9	The nightlights on all 3 stopped working less than 6 months after I started using them. I called the company and was told that if the nightlights are out then the product isn't working. They advertise that these will last for 3-5 years. Did the product work when the nightlights were working? I don't know because I was still trapping mice in the traps that were nearby!
Review 10	Excellent. I put 2 in each room. I haven't seen a trace since.

Table 11: Example summary generated by RANDSEL, where reviews are randomly sampled in training and testing. Extrinsic hallucinations are marked in red, intrinsic in orange. When random review subsets are used, the decoder is forced to invent 'novel' content in training leading to hallucinations in test time.