Opinion Summarization using High-Quality Synthetic Dataset

Rajwinder Mahal
rsm2207@columbia.edu

Abstract

Opinion summarization is the task of generating a summary of the opinions expressed in a set of documents. This is a challenging task as it requires the model to understand the sentiment of the input documents and to identify the key information. The task is practically important and has attracted a lot of attention. However, due to the high cost of annotating data, there is a lack of large summary datasets for supervised models. Instead, the task has been traditionally approached with extractive methods that extracts key information from the text in an unsupervised way. More recent unsupervised methods have shown promising results for abstractive summaries, however, these models fail to capture their essential properties. To address these problems, we first used large language models to generate (review, summary) pairs to create a labeled dataset, and then we use this dataset to fine-tune a much smaller model. Our simple approach enables the model to generate high-quality summaries by just using a small dataset of high-quality synthetic/generated dataset. Experiments on Yelp reviews show that this approach results in fluent and coherent summaries reflecting common opinions.

1 Introduction

Websites such as Amazon and Yelp allow customers to leave reviews to share their experiences, offer recommendations, and express their satisfaction or dissatisfaction with a product or service. As writing reviews is becoming a common practice, most people cannot imagine making a purchasing decision without first understanding how other customers who spent their money, got the product, and used it feel about it. This proliferation of online reviews has accelerated research on opinion summarization due to its potential for not only influencing the decisions of potential customers but also for various business intelligence applications such as creating reports, analyzing user behavior, optimizing search queries, and personalized recommendations (Hu and Liu, 2004; Medhat et al., 2014; Angelidis and Lapata, 2018).

Although significant progress has been made in supervised summarization tasks (Rush et al., 2015; Chopra et al., 2016; Liu and Lapata, 2019), these deep learning methods rely on large amounts of annotated data which is either not available or is very expensive to produce. This lack of sufficiently large annotated datasets led to various experiments to explore unsupervised methods for opinion summarization. Recent unsupervised models (Chu and Liu, 2019; Amplayo and Lapata, 2020; Bražinskas et al., 2020b) use a multi-step approach where they first extracts key information from the reviews and then generates the summary by conditioning on the extracted data. However, these models are trained on smaller, toy datasets and are never exposed to actual summaries. As a result, the generated summaries mimic the informal style of reviews and contain unwanted details. These limitations are addressed by more recent few-shot, supervised methods such as FewSum (Bražinskas et al., 2020a) that consists of a transformer-based generator followed by a plug-in network that switches the generator into a summarizer. It is trained on a small dataset and is able to generalize to new domains and target entities with only a few examples.

The recent surge in performance of large language models (LLMS) has led to a desire to use them for opinion summarization. (Bhaskar et al., 2023) explores the use of GPT-3.5 to summarize a large collection of user reviews in a prompted fashion. It uses pipeline methods to summarize text using different approaches such as recursive summarization and aspect-oriented extractive methods. Based on this idea, a more successful approach is to generate a synthetic dataset (Conover et al., 2023; Taori et al., 2023), where (review, summary) pairs are constructed from a review dataset to enable supervised training. This approach solves the challenges associated with annotating a large dataset of reviews.

In this paper, we use prompting with PaLM 2 (Anil et al., 2023) to generate a high-quality labeled dataset of review and summary pairs given the yelp reviews (Yelp, 2023). Generating dataset in this setting is not straightforward, as the combined length of the reviews may exceed the model's maximum input length. Furthermore, we find many unpopular opinions are ignored by the model. For example, if the review is positive overall and there's only one attribute that's reviewed negatively, it is not guaranteed that this unpopular opinion will be present in the output summary. To mitigate these issues, we explore different data generation approaches such as conditional generation given the business attributes, using business rating as a conditional variable to make model learn whether to put more focus on positive or negative opinions, and sentiment based approach to summarize different opinions separately and then aggregating them to generate the final summary.

We further utilize a pre-trained BART-base (Lewis et al., 2019) model and fine-tune four different models to experiment with different fine-tuning approaches. Finally we show that our model can generate high quality summaries that are on par with the summaries generated by a much larger model. Our contributions can be summarized as follows:

- we introduce a simple synthetic dataset-based approach to generate opinion summaries;
- we demonstrate that the approach substantially outperforms the previous methods, both when measured with automatic metrics and in human evaluation;
- we provide a dataset of abstractive summaries for Yelp reviews.

2 Related Work

Opinion summarization is a challenging task that aims to generate a concise summary of the opinions expressed in a text. It has a wide range of applications in e-commerce, including product recommendations, market research, and customer support (Hu and Liu, 2004; Moussa et al., 2018; Bhatia, 2021). Earlier work on opinion summarization has focused on extractive methods (Hu and Liu, 2006; Kim et al., 2011; Angelidis and Lapata, 2018), which involve identifying the most important sentences from a text and combining them into a summary. Some of the most widely used extractive summarization techniques are based on Latent Semantic Analysis (LSA) (Gong and Liu, 2001; Dumais, 2004) and Bayesian Topic Modeling (Daumé and Marcu, 2006; Haghighi and Vanderwende, 2009). LSA creates a matrix that represents the importance of words in sentences, and uses singular value decomposition to select the most relevant sentences. Bayesian Topic Modeling uses a generative model to represent documents as mixtures of latent topics, where a topic is a probability distribution over words. However, these extractive methods are not well suited for generating long summaries that are coherent and fluent, with the exception of few graph-based methods such as LexRank (Erkan and Radev, 2004).

Abstractive opinion summarization is an emerging branch that generates a new summary text by understanding and processing the original text (Ganesan et al., 2010; Di Fabbrizio et al., 2014). Unlike extractive summarization, abstractive summarization can produce more informative and human-readable summaries. More recent work has seen the effective application of sequence-to-sequence models to generate document representations which are then used to generate a new summary (Rush et al., 2015; Chopra et al., 2016; Liu and Lapata, 2019). However, due to the absence of opinion summaries and the difficulty of annotating a large dataset, the recent models (Chu and Liu, 2019; Amplayo and Lapata, 2020; Bražinskas et al., 2020b) perform opinion summarization in an unsupervised way. These models typically generate opinion summaries in a two-stage process which first extracts key information from the reviews in an extractive way, and then generates the summary by conditioning on the extracted data. For example, CopyCat (Bražinskas et al., 2020b) is an unsupervised abstractive method that models sentences as observations of hierarchical continuous latent representations to model entity opinions. Although these models have shown promising results, these are mostly done on toy datasets and are never exposed to actual summaries.

3 Data

Our experiments are conducted on Yelp dataset (Yelp, 2023), which is a subset of Yelp's businesses, reviews, and user data. This dataset consists of over

One should never expect perfection from a fast food chain. It's understandable the product never looks like the commercials. But when 35 of the soft tacos you get have two pieces of lettuce, and one third of the length of the taco is barely covered in meet, with the other part meatless, that's not ok. They gave us all of the wrong sauces, and the crunch part of the cheesy Gordita crunch, was as crumbly as the little tiny chips left over in the stale bag of three week old Doritos from the bottom of the chip drawer...
We had Mark at our new home for an inspection. He claimed most everything was ok, and found nothing major. We just moved in and found we have NO COLD WATER TO THE TUB AND OUR DRYER VENT IS FILLED WITH LINT- fire hazard. I called Mark and as I expected he said everything was working...

Table 1: Sample reviews from the Yelp dataset.

6.9 million reviews across 150k businesses, including restaurants, retail shops, service providers. The data is formatted as JSON object with various attributes related to business entity and its reviews. For the business entity, there are common fields such as business name, address, ratings, number of reviews, business categories, etc. Reviews consist of review text along with user_id that wrote the review and the business_id the review is written for.

	Count	Mean	Min	Max
No. of businesses	150346			
Ratings	150,243	3.59	1.0	5.0
Reviews per business		44.89	5.0	7568.0

Table 2: Statistics on Yelp Businesses.

	Count	Mean	Min	Max
No. of reviews	6,990,280			
Ratings	6,990,280	3.75	1.0	5.0
Review length		567.76	1.0	5000.0
Review # words		104.78	1.0	1070.0

Table 3: Statistics on Yelp reviews.

Yelp reviews contain a lot of personal information and irrelevant details which one may find unnecessary in a summary (Bražinskas et al., 2020b) For example, a review for a restaurant that customer visited for a birthday party includes details related to the party that are irrelevant when summarizing reviews for a restaurant. Therefore, our models need to distill important information in reviews while abstracting away from details such as mentions of specific dates or occasions upon which customers visited a restaurant.

These Yelp reviews are used as an input to PaLM 2 API to generate a labeled dataset of review-summary pairs. This generated dataset is then used to fine-tune out proposed models. Details of this generated dataset are included in the next section.

To evaluate our generated dataset, prompts, and fine-tuned models, we used Yelp dataset released by Bražinskas et al. (2020a). The dataset contains 300 human-written summaries for 100 Yelp businesses. These summaries are generated by Amazon Mechanical Turk (AMT) workers. Three AMT workers summarized each business, generating 3 different human-written summaries for each business.

	Count	Mean	Min	Max
No. of reviews	800			
Review # words	800	49.49	25	66
Summary # words	300	50.85	25	88

Table 4: Statistics on Yelp reviews from the evaluation dataset released by Bražinskas et al. (2020a).

4 Methods

4.1 Task Formulation

The main objective of our model is to generate a review summary R_{sum} given multiple customer reviews $R_{0...N}$ of a business entity. The summary R_{sum} should (i) capture positive, negative, and neutral opinions, (ii) minimize the inclusion of irrelevant details such as personal information from the input reviews, and (iii) include key business aspects such as the quality of the food, the service, the atmosphere, and the price.

4.2 Data Generation Approach

Recently, the escalating capabilities of large language models (LLMs) have shown promising results for generating datasets to fine-tune smaller models (Conover et al., 2023; Taori et al., 2023). Based on this idea, the first step of our approach is to generate a dataset of (review, summary) pairs using the Yelp dataset. We conducted multiple experiments to evaluate different prompting approaches:

Simple Prompting Given a set of review $R_{1...N}$ for a business entity S_e, we prompt PaLM 2 to output the target summary S_e.

Controlled Generation using Business Attributes Given a set of review $R_{1...N}$ for a business entity S_e, we first use PaLM 2 to generate business attributes $A_{e,1...N}$. We modified our prompt to condition the generated summary S_e on the top 10 business attributes $A_{e,1...10}$.

Sentiment based Summary Given a set of review $R_{1...N}$ for a business entity S_e, we first prompt PaLM 2 to output the target summaries S_{pos} and S_{neg}, where *pos* and *neg* refer to positive and negative opinions, respectively. We use these summaries as an input to the PaLM 2 to generate the final summary S_e.

4.3 Our Models

We designed our fine-tuning approach based on the approaches used by Lee, Bang, Yu, Madotto, and Fung (2022); Prabhumoye, Patwary, Shoeybi, and Catanzaro (2023). We fined-tuned four BART base (Lewis et al., 2019) models to evaluate different approaches.

4.3.1 Vanilla Model

We fine-tuned our first model to simply summarize reviews. The input X to our BART encoder is formatted as follows:

$$R_{e,1...N}$$

where $R_e, 1$ is the first review of the business entity e.

The generated target Y of our BART's autoregressive decoder is formatted as follows:

$$S_e$$

where S_e is the generated summary of the business entity e.

4.3.2 Controlled Model: Business Attributes

Some recent methods (Amplayo et al., 2020; Amplayo and Lapata, 2021) have used business attributes to condition the generated summary on these attributes. So, we experimented with this approach to generate high-quality summaries using the generated dataset.

The input X to our BART encoder is formatted as follows:

$$businessattributes : A_{1...N}[SEP]$$

$$R_{e,1...N}$$

The generated target Y of our BART decoder is formatted as follows:

$$S_e$$

4.3.3 Controlled Model: Business Rating

Drawing inspiration from Prabhumoye et al. (2023), who explored adding toxicity score during pretraining to significantly reduce model toxicity, we explored a similar approach for summarization by incorporating business rating.

Approach 1 The input X to our BART encoder is formatted as follows:

$$rating : r_e[SEP]$$

$$R_{e,1...N}$$

where r_e is the numeric rating of business entity e.

Approach 2 For this approach, we converted the numeric rating into a text prompt using the mean rating score as the cutoff.

The input X to our BART encoder is formatted as follows:

$$rating : businesshashighratings.[SEP]$$

$$R_{e,1...N}$$

OR

$$rating : businesshaslowratings.[SEP]$$

$$R_{e,1...N}$$

5 Experiments

In this section, we describe in more details our process for generating data, fine-tuning our models, selected baselines, and evaluation methodology.

5.1 Experimental Details

We used a standard Transformer encoder-decoder (Vaswani et al., 2023) model, pre-trained BART base (Lewis et al., 2019), consisting of 140M parameters with 6 encoder and decoder layers, 1024 hidden states, and 16 attention heads.

We used HuggingFace Transformers library for fine-tuning our models with a learning rate of $5e-5$ for 600 steps. We used Adam with weight decay (Loshchilov and Hutter, 2019) for parameter optimization. We used 2 Tesla L4 GPUs for each experiment, each with a batch size of 4 and 4 gradient accumulation steps; effective batch size of 32. We performed summary generation with beam search of size 3. It is common practice to set the number of data loading workers to the number of CPU cores. In our case, we used 4 data loading workers per GPU process to reduced GPU idle time.

5.2 Dataset and Data Generation

Our data preprocessing include filtering out low rating businesses, remove excessively long or short reviews, etc. We filtered out all businesses with low ratings or low review counts to reduce noise, outliers, or uninformative content. We also filtered out excessively long or short reviews, including ones with low ratings. Statistics of original and processed datasets are in Tables 2, 3 and 5.

Due to limited resources, we created a mini-dataset for our experiments. We randomly selected businesses with a minimum of 15 reviews and 3.0 rating. For each business, we randomly selected 8 reviews with a minimum rating of 3.0, minimum word length of 42 and maximum word length of 149. Our training set has 6,729 businesses with a total of 53,832 reviews. Validation set has 748 businesses with a total of 5,984 reviews. For our test set, we used Yelp dataset released by (Bražinskas et al., 2020a), which contains 300 human-written summaries for 100 Yelp businesses.

We used PaLM 2 API (text-unicorn@001 model) to generate summaries. For each experiment, we experiment with different prompts and evaluated summaries both manually and using automatic metrics to find the best possible prompts. Initially, we experimented with BERTopic (Grootendorst, 2022) and KeyBERT (Grootendorst, 2020), but it didn't perform well. So, we decided to use PaLM 2 API to generate business attributes.

	Count	Mean	Min	Max
No. of businesses	37,269			
Business ratings	37,269	4.01	3.0	5.0
Review # words	2,327,562	82.28	42.0	149.0

Table 5: Yelp dataset processed

	Count	Mean	Min	Max
No. of businesses	7,477			
Review # words	59,816	82.51	42.0	149.0
Generated summary # words	7,477	72.77	59.0	117.0

Table 6: Yelp generated dataset

5.3 Baseline Models

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 auto-encoder trained to reconstruct reviews of a product.

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

FEWSUM (Bražinskas et al., 2020a) is a few-shot framework where lexical features are used to differentiate between customer reviews and summaries. In the fine-tuning phase, features leading to generation of summaries are searched.

5.4 Evaluation Metric

Both automatic and manual strategies are used throughout experiments to evaluate the perfor-

	R_1	R_2	R_L	Precision	Recall	f1	BLEU	METEOR
Simple Prompting	0.3473	0.0876	0.2173	0.8714	0.8779	0.9745	0.047	0.2962
Business Attributes Based	0.3599	0.0887	0.2229	0.8845	0.887	0.8856	0.048	0.2938
Sentiment Based	0.3358	0.079	0.2077	0.8765	0.8848	0.8806	0.0412	0.2798

Table 7: Evaluation results from the data generation experiments on the evaluation dataset.

	R_1	R_2	R_L	Precision	Recall	f1	BLEU	METEOR
Our vanilla model	0.3665	0.0906	0.229	0.8874	0.8844	0.8858	0.0579	0.2747
Our Controlled business attributes	0.3614	0.0855	0.2294	0.888	0.8839	0.8858	0.0548	0.2723
Our Controlled numeric ratings	0.3662	0.0917	0.2286	0.8868	0.8839	0.8853	0.0585	0.2746
Our Controlled ratings as prompt	0.368	0.092	0.2307	0.8867	0.884	0.8853	0.0595	0.2752
FewSum	0.3729	0.0992	0.2276					
Copycat	0.2812	0.0589	0.1832					
MeanSum	0.2750	0.0354	0.1609					
LexRank	0.2696	0.0493	0.1613					

Table 8: Evaluation scores on the Yelp test dataset with human-written gold summaries.

mance of fine-tuned model. We used industry-standard Recall-Oriented Understudy for Gisting Evaluation (ROGUE) (Lin, 2004) and Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) scores to evaluate the performance of fine-tuned model. ROUGE measures the quality of a generated summary by capturing relevant content from the reference text, even if the wording or phrasing differs. BLEU measures the correctness and exact matches between n-grams in the generated and reference text, and often doesn't account well for synonyms or variations in wording. To complement the shortcomings of ROUGE, we also used BERTScore (Zhang et al., 2020) to measure the similarity between generated and reference summaries at the contextual level, rather than just relying on n-gram overlap, longest common sequences, or weighted word overlap. We also used Metric for Evaluation of Translation with Explicit Ordering (METEOR) (Banerjee and Lavie, 2005) score to measure the quality of generated text based on the alignment between the generated text and the reference text.

6 Results

Data Generation Table 7 shows our data generation results on the Yelp test dataset released by Bražinskas et al. (2020a). Our experiments shows that the multi-step prompting approach yields the best results. This approach first prompt the PaLM 2 model to generate a set of business attributes for a business based on a given set of reviews. Then, we take the top 10 business attributes to prompt the model again to generate summary conditioned on these attributes. We also noticed that multi-step sentiment-based approach performed worse than the simple prompting approach. This might be due to model generating positive and negative opinions when asked to do so even if not both are present in the reviews.

Fine-tuned models Table 8 shows the evaluation results of our fine-tuned models. Our controlled model with rating as prompt prefixed to input reviews performed the best. We also noticed that the vanilla model performed better than the controlled model on business attributes or numeric rating. Our intuition was that the controlled model on business attributes would result in best performance as data generation using business attributes yielded the best results. Moreover, we can see that our model performed better than the all baselines except FewSum. FewSum also uses language model for few-shot learning to generate summaries. So, this validates our idea of using large language models to gen-

Reviews	Best breakfast in Akron. They care about quality and it shows in the food. The bar is designed like an island in your kitchen, makes you feel like you're at home. Wait staff is very professional and treat you like family. Owner is local and is very hands on; which shows in the food. Great place for lunch too!...Very busy on weekends, yet there is never too long of a wait. Service is great and the portions are very generous, especially the pancakes! (My favorite). The new interior is very nice and adds great atmosphere!...What a great place! Food is amazing and it's not just your ordinary breakfast or lunch spot. The food is unique and delicious. One of my favorite is the red eye hash! Fresh orange juice or a bloody mary, either way both delicious. The service is always great and the atmosphere is good....One of our favorite breakfast spots. There's often a wait on the weekends but we've never waited more than 10 minutes. The host and wait staff and always friendly and accommodating and the food is consistently wonderful. I recommend the eggs Benedict with crabmeat or the red eye hash!...A very good breakfast spot that has it's own take on popular dishes. The potatoes are especially tasty, though everyone at our table enjoyed their meals which ranged from pancakes, eggs and french toast. I also like the Akron themed pictures on the wall.
keywords	food, service, atmosphere, wait time, price
Gold summ1.	This restaurant has consistently good food and service. It is an especially popular place for breakfast, though they serve a tasty lunch as well. The atmosphere inside is positive and the staff are always friendly. Expect a short wait on the weekends, as it can become overcrowded.
Gold summ2.	Really great restaurant for a nice breakfast! Fantastic and unique dishes that never fails to amaze customers, friendly and efficient staff, generous portions and great atmosphere. Excellent menu with a wide variety. Management is quality-minded. Overall a highly recommended place.
Gold summ3.	This restaurant is often very busy on weekends, but even so there usually isn't much of a wait. The staff is very friendly and provide great service. The food is a bit unique, but all of it is very good, particularly the eggs and pancakes. They specialize in breakfast, but also offer sandwiches for lunch. The portions are large for the price they charge. This place is highly recommended.
PaLM 2 summ.	This restaurant is a local favorite for breakfast and lunch. The food is delicious and unique, and the portions are generous. The service is friendly and attentive, and the atmosphere is casual and inviting. There is often a wait on the weekends, but it is worth it. The prices are reasonable.
Our summ.	Omlets is a great breakfast spot in Akron. The food is delicious and the portions are generous. The staff is friendly and accommodating, and the atmosphere is casual and relaxed. The prices are reasonable, and there is often a wait on weekends, but it's worth it for the delicious food.

Table 9: Sample reviews along with gold, PaLM 2 generated, and our best model generated summaries. The sample is from the Yelp test dataset.

erate summaries as both models performed well. We also noticed a larger gain compared to other unsupervised baselines.

We also manually inspected the generated summaries and compared it to the both human-written gold summaries and summaries generated by PaLM 2. Table 9 shows the generated summaries using the PaLM 2 and our best fine-tuned model. We found that our summaries are even more fluent and coherent than the human-written summaries. Most of the human-written summaries include unwanted details but our generated summaries are concise and to the point.

7 Error Analysis

Our experiments have produced good results, however, we found few examples were our approach didn't perform well. We believe mitigating such cases will help us further improve our model. For example, Table 9 shows summary generated by our model. We can see that it named the restaurant "Omlets" in the generated summary, however, it is not mention anywhere that the restaurant name is "Omlets". We think this might be because some reviews mentioned that the restaurant owner is local and is very friendly. So, we think our model mistakenly thought that the name of the restaurant is "Omlets" (confusion b/w "Omlets" and "Owner").

During our data generation stage, we also notice that sometimes generated business attributes are not formatted as expected. We tried different prompts and data cleaning steps but our generated dataset still contains some examples formatted incorrectly. Another error we found was that some generated summaries were too short which might propogate the error from data generation stage to our fine-tuned models.

We also inspected human written summaries that are used by Bražinskas, Lapata, and Titov (2020b) and Bražinskas, Lapata, and Titov (2020a) and found that these summaries are very different from the summaries generated by our model as well as summaries generated by PaLM 2. These human-written summaries, as seen in Table 9, contain some unwanted details such as specific food items. Additionally, we have 3 human-written summaries per business, however, these summaries differ a lot from each other. We think that the automatic metrics are not able to fully capture the results of our model as we compared our model's generated results with these human-written summaries which looks inferior as compared to summaries generated by our model or PaLM 2.

8 Conclusions, Limitations, and Future Work

In this paper, we presented a new approach to abstractive summarization of Yelp reviews, which only uses a small, high-quality synthetic dataset to fine-tune a model that produces fluent and coherent summaries reflecting common opinions. Our fine-tuned model performs much better than the unsupervised methods and is on par with the FewSum, which uses few-shot learning approach.

We think that there is a need for further experimentation to understand the fundamental relationship between business ratings and generated summaries. Although our model performed well, we think this behavior might change if we include low rating businesses or reviews with low ratings in our dataset. Furthermore, our experiments only use 8 reviews per business, however, the dataset contains hundreds and even thousands of reviews for some of the businesses and it would be an interesting experiment to check how our model performs when it is given a large number of reviews. Additionally, there are limitations on how much data we can input to a model, so we plan to slightly modify our approach to use an iterative process to summarize large number of reviews Bhaskar, Fabbri, and Durrett (2023).

Furthermore, it is important to do proper human evaluations as we have seen that some of the human-written summaries were not as good as expected. This mean that the automatic metrics were not able to capture the full capability of our model. We manually reviewed some of the generated summaries, but due to limited time and resources, we couldn't do a proper human evaluation. Overall, our approach substantially outperforms the previous methods, both when measured with automatic metrics and manual review.

References

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2020. Unsupervised opinion summarization with content planning.

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. 2021. Informative and controllable opinion summarization.

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.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Adithya Bhaskar, Alexander R. Fabbri, and Greg Durrett. 2023. Prompted opinion summarization with gpt-3.5.

Surbhi Bhatia. 2021. A comparative study of opinion summarization techniques. *IEEE Transactions on Computational Social Systems*, 8(1):110–117.

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.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, San Diego, California. Association for Computational Linguistics.

Eric Chu and Peter Liu. 2019. MeanSum: A neural model for unsupervised multi-document abstractive summarization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1223–1232. PMLR.

Mike Conover, Matt Hayes, Ankit Mathur, Xiangrui Meng, Jianwei Xie, Jun Wan, Ali Ghodsi, Patrick Wendell, and Matei Zaharia. 2023. Hello dolly: Democratizing the magic of chatgpt with open models.

Hal Daumé and Daniel Marcu. 2006. Bayesian query-focused summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, page 305–312, USA. Association for Computational Linguistics.

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, Philadelphia, Pennsylvania, U.S.A. Association for Computational Linguistics.

Susan T. Dumais. 2004. Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1):188–230.

G. Erkan and D. R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

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.

Yihong Gong and Xin Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, page 19–25, New York, NY, USA. Association for Computing Machinery.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure.

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

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*, KDD '04, page 168–177, New York, NY, USA. Association for Computing Machinery.

Minqing Hu and Bing Liu. 2006. Opinion extraction and summarization on the web. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, page 1621–1624. AAAI Press.

Hyun Duk Kim, Kavita Ganesan, Parikshit Sondhi, and ChengXiang Zhai. 2011. Comprehensive review of opinion summarization.

Nayeon Lee, Yejin Bang, Tiezheng Yu, Andrea Madotto, and Pascale Fung. 2022. Neus: Neutral multi-news summarization for mitigating framing bias.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

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.

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

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

Mohammed Elsaid Moussa, Ensaf Hussein Mohamed, and Mohamed Hassan Haggag. 2018. A survey on opinion summarization techniques for social media. *Future Computing and Informatics Journal*, 3(1):82–109.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Shrimai Prabhumoye, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Adding instructions during pretraining: Effective way of controlling toxicity in language models.

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.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Yelp. 2023. Yelp open dataset: An all-purpose dataset for learning. `https://www.yelp.com/dataset`. Accessed: 2023-10-03.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.