OpinioRAG

Towards Generating User-Centric Opinion Highlights from Large-scale Online Reviews

Mir Tafseer Nayeem and Davood Rafiei

University of Alberta

Motivation

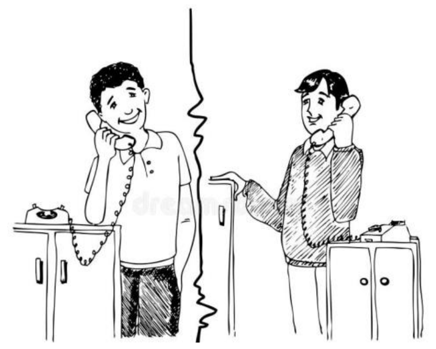
What others think has always been an "important" piece of information.

"Which hotel should I book?"
"Which professors to work for?"

"So whom shall I ask?"

Pre Web

- Friends and relatives
- Person with knowledge
- Customer reports

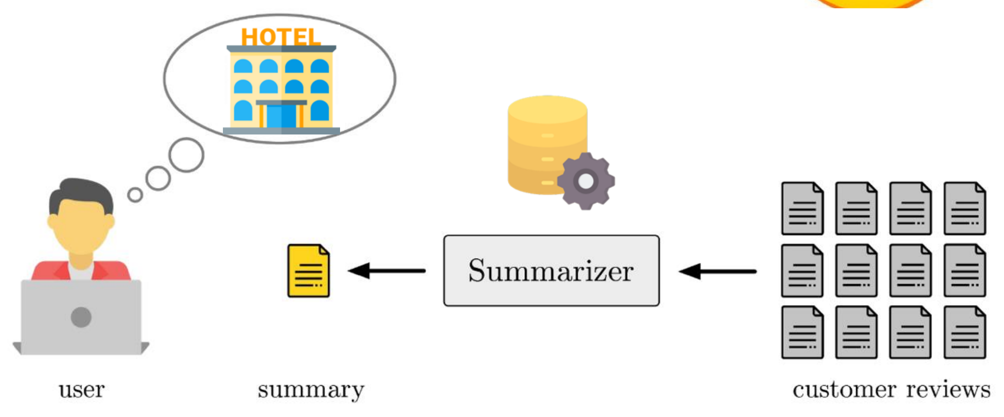


Post Web

- E-commerce (Amazon)
- Review sites (CNET)
- Discussion forums

Challenges

- **Content Volume**
  - Sheer volume (*"too much"*) of reviews → **"information overload"**
  - Users skim a subset of reviews → **suboptimal decisions**
- **Absence of Explicit Structure**
- **Noisy and Repetitive**
- **Stylistically Diverse Inputs**



Research Gaps

- **Short-form inputs** (e.g., mostly 10 reviews), inadequate for real-world scenarios.
- Relatively **mid challenges** for modern LLMs.
- **One-size-fits-all** summaries, fail to cater personalized user needs
  - *"room cleanliness", "public transport", "fitness facilities",* or *"pet-friendly policies"*
- **Generate generic**, paragraph-style summaries.
  - Not useful for informed decision making

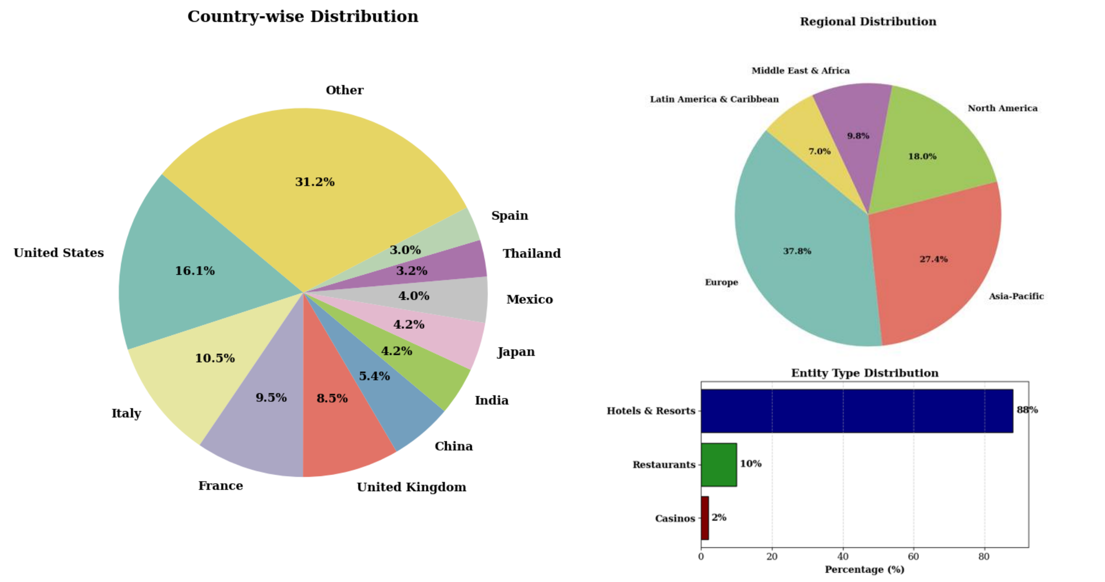
OpinioBank

Goal: to advance user-centric opinion summarization over large-scale (>100K), noisy, and diverse inputs.

- Data Sources
  - **Source:** TripAdvisor
  - **Target:** Oyster
- Entity Linking & Crwaling
- Manual Query Annotation
- Review Alignment Verification
- Metadata Integration

tripadvisor · Oyster

Unbiased Reviews · In-Depth · Travel Expertise



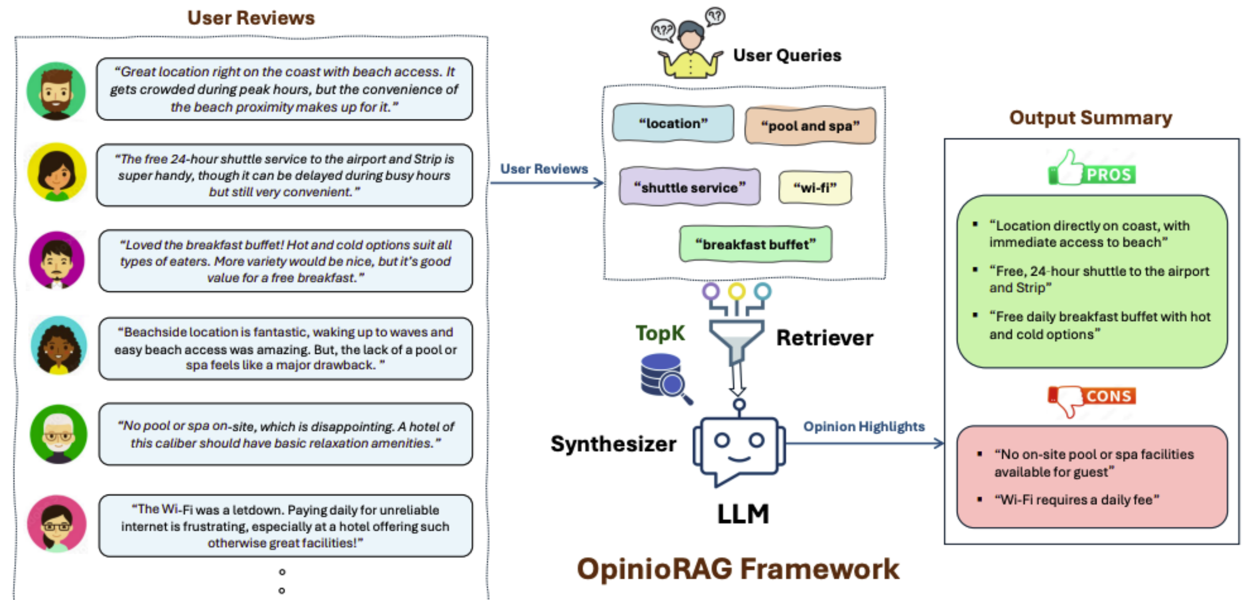
Datasets	Domains	#Entities	#Queries	Unique #Queries	#Revs	#Sents	#Tokens	Book Len?	Expert Pairs	Meta data	P & C
MeanSum (2019)	Businesses	200	✗	✗	8	41.1	561.01	✗	✗	✗	✗
CopyCat (2020b)	Products	60	✗	✗	8	30.38	463.62	✗	✗	✗	✗
FewSum (2020a)	Businesses	60	✗	✗	8	29.85	457.05	✗	✗	✗	✗
OpoSum+ (2021)	Products	60	240	4	10	71.8	1,194.0	✗	✗	✗	✗
SPACE (2021)	Hotels	50	350	7	100	910.58	16,770.18	✗	✗	✗	✗
AmaSum (2021)	Products	3,166	✗	✗	322.31	1,057.3	15,614.71	✗	✓	✗	✓
ProSum (2024)	Restaurants	500	✗	✗	6.70	71.34	1,236.38	✗	✓	✗	✗
OpinioBank (*ours*)	Hotels	500	5,975	1,456	1.5K	10.5K	207K	✓	✓	✓	✓

Table 1: Comparison of our **OpinioBank** dataset with existing alternatives, focusing on long-form inputs (over 100K tokens) and user queries. #Entities denotes dataset size, #Queries refers to query count, #Revs indicates average reviews per entity, #Sents represents average sentences, and #Tokens indicates average tokens (using GPT-4o tokenizer) per entity. P & C stands for PROS & CONS. Other availabilities are indicated using ✓and ✗.

OpinioRAG: Two Stages

Attributability and scalability of extractive RAG methods and Coherence and fluency of LLMs.

- **Scalable and training-free** solution for generating user-centric opinion highlights from long-form inputs.
- **Structure outputs** around specific user queries.



OpinioRAG Framework

Retrieval Stage

Extract the most relevant ones as evidence and reduces clutter by filtering key evidence before generation.

1. Lexical (BM25)
2. Semantic (Dense)

Retrieval Function · User Review Corpus

$$S_Q = \text{Top-K}\,(\mathcal{R}(Q, D))$$

TopK Sentence · Query Set

Synthesizer Stage

- The retrieved evidence is then utilized to generate query-specific highlights using LLMs.
- Structured outputs in a predefined JSON format while adhering to the desired key-point style.

TopK Sentence · User Query · Instruction

$$\text{Highlight}(Q) = \text{LLM}(Q, S_Q, \mathcal{C}, \mathcal{E}, \mathcal{P})$$

Stylistic Guidelines and Exemplars

Controllability
Scalability
Modularity
Verifiability

RAG Verification

- Novel verification metrics evidence-highlight level.
- Decompose sentences into structured components.
- Fine-grained assessment of factual alignment.

Retrieved Evidence: $\{(a_R^i, o_R^i, s_R^i)\}_{i=1}^n$, Generated Highlight: (a_G, o_G, s_G)

Aspect Relevance (AR)

$$a^* = \arg\max_{a \in \mathcal{A}} \text{freq}(a, R)$$

$$AR = 1\,(a^* = a_G)$$

Sentiment Factuality (SF)

$$s^* = \arg\max_{s \in \{-1,1\}} \text{freq}(s, R|a)$$

$$SF = 1\,(s_G = s^*)$$

Opinion Faithfulness (OF)

- Direct match is a score of 1 and indirect matches are computed using a semantic similarity function.

Evaluation

1. Lexical (ROUGE)
2. Semantic (BERTScore)
3. LLM-as-a-Judge

Baselines		PROS Scores R1	R2	RL	BS	CONS Scores R1	R2	RL	BS
Random		16.28	1.39	9.53	53.06	10.09	0.50	7.22	51.77
Extractive Oracle		50.51	17.66	40.96	71.25	39.89	10.77	33.09	66.61
TextRank		16.56	2.05	9.64	54.57	10.17	0.61	7.11	51.89
LexRank		16.68	1.81	9.74	54.90	10.64	0.59	7.19	52.08
Model IDs	**CL**	**Long-context LLMs**							
GPT-4o-mini	128K	29.97	5.76	17.21	64.95	18.97	2.70	12.22	59.94
Claude-3.5-haiku	128K	32.70	7.03	19.30	**67.37**	20.07	3.03	13.44	61.41
Gemini-2.0-flash	1M	30.62	5.75	17.87	65.45	20.81	3.73	13.70	60.92
Models/Ablations	**Acc.**	**OpinioRAG (*ours*)**							
BM25 (K=10)		30.80	5.90	22.05	60.87	27.83	5.57	22.13	60.79
├ GPT-4o-mini	🔒	35.92	7.98	25.94	64.84	30.59	6.90	24.42	64.26
├ Gemini-2.0-flash	🔒	33.95	6.65	24.01	62.54	29.45	6.38	22.67	62.71
├ Claude-3.5-haiku	🔒	35.89	8.52	26.65	66.53	29.08	6.12	23.48	63.66
├ Gemma-2-9B	🔓	34.77	7.18	26.45	64.75	**33.05**	<u>8.08</u>	**27.34**	<u>65.62</u>
├ Mistral-7B	🔓	36.30	8.43	<u>27.07</u>	66.28	32.47	7.40	26.17	64.38
├ Llama-3.1-8B	🔓	**37.51**	**9.13**	**27.41**	66.62	<u>32.61</u>	8.06	25.79	64.79
Dense (K=10)		28.86	4.99	20.77	61.91	25.37	4.64	20.11	60.82
├ GPT-4o-mini	🔒	35.69	7.66	25.96	65.55	29.48	6.70	23.52	64.19
├ Gemini-2.0-flash	🔒	33.97	6.58	24.16	63.42	28.74	5.90	22.23	62.94
├ Claude-3.5-haiku	🔒	35.27	8.05	26.19	66.76	27.52	5.18	22.37	63.58
├ Gemma-2-9B	🔓	34.45	6.56	25.86	65.14	32.31	**8.32**	<u>26.84</u>	**65.81**
├ Mistral-7B	🔓	36.33	8.20	26.97	<u>67.19</u>	31.38	7.09	24.94	64.24
├ Llama-3.1-8B	🔓	<u>36.86</u>	<u>8.49</u>	26.85	66.88	31.56	6.97	24.54	64.50

Table 2: Performance comparison of various models and retrieval methods (TopK = 10) in the OpinioRAG framework against baselines and long-context LLMs. The results are evaluated using lexical-based metrics (R1, R2, RL) and the embedding-based metric BERTScore (BS) for 'PROS' and 'CONS'. The icons 🔓 and 🔒 indicate open-source and closed-source models. **Bold** and <u>underlined</u> values denote the best and second-best results for each metric.

RAG Verification Assessment

Models	TopK (K = 5) BM25 AR	SF	OF	Dense AR	SF	OF	TopK (K = 10) BM25 AR	SF	OF	Dense AR	SF	OF
GPT-4o-mini	75.30	88.63	76.75	73.91	88.76	77.55	76.62	89.16	78.20	74.71	**89.75**	79.65
Gemini-2.0-flash	**79.24**	87.90	80.13	77.18	87.87	80.52	<u>78.98</u>	86.93	<u>82.50</u>	78.30	87.91	**82.75**
Claude-3.5-haiku	76.43	88.40	71.79	75.31	86.91	71.98	77.22	86.82	74.22	75.49	87.58	74.80
Gemma-2-9B	76.78	88.46	78.42	75.83	88.03	79.32	77.89	87.71	81.47	77.15	<u>89.32</u>	82.14
Mistral-7B	75.89	86.30	78.65	74.65	86.68	78.46	77.31	87.04	80.76	74.14	87.52	81.15
Llama-3.1-8B	77.65	87.45	78.34	73.99	87.21	79.80	78.82	87.40	81.11	74.28	88.91	82.35
AVG.	76.88	87.86	77.35	75.15	87.58	77.94	77.81	87.51	79.71	75.68	88.50	80.47

Table 3: Comparison of Aspect Relevance (AR), Sentiment Factuality (SF), and Opinion Faithfulness (OF) across various models using BM25 and Dense retrieval methods for TopK = 5 and TopK = 10. Results indicate that increasing TopK generally improves performance. BM25 is more effective for AR, while Dense retrieval performs better for SF and OF.

LLM-as-a-Judge Evaluation

Evaluation Criteria

Aspect Relevance (AR)	Does the system summary cover the same topics or facets as the expert summary?
Non-Redundancy (NR)	Are aspects mentioned only once? Are key points repeated or paraphrased redundantly?
Sentiment Agreement (SA)	Is the tone (positive or negative) about aspects consistent between the summaries?
Opinion Faithfulness (OF)	Are the factual or evaluative claims in the system summary grounded in the expert summary?
Overall Usefulness (OU)	Would the system summary help a potential customer make a reasonable decision?

Figure 2: LLM-as-a-Judge evaluation criteria used to assess the quality of the summaries.

Key Insights

- Long-context LLMs struggle to retrieve and synthesize.
- BM25 outperform Dense.
- Extracting critical drawbacks ('CONS') is challenging.
- Oracle indicates substantial room for improvement.
- TopK = 5 to TopK = 10 consistently improves.
- BM25 excels in (AR) and Dense in (SF, OF).

Model	Type	AR	NR	SA	OF	OU
BM25 (K=10)						
Gemma-2-9B	🔓	3.14	3.81	2.93	2.88	3.11
Mistral-7B	🔓	3.25	3.72	3.08	2.90	3.19
Llama-3.1-8B	🔓	3.26	<u>3.85</u>	2.98	2.86	3.19
GPT-4o-mini	🔒	3.17	3.57	2.90	2.86	3.12
Gemini-2.0-flash	🔒	3.29	3.46	2.93	2.91	3.18
Claude-3.5-haiku	🔒	3.24	3.70	3.04	2.92	3.19
Dense (K=10)						
Gemma-2-9B	🔓	3.25	3.60	3.14	2.96	3.18
Mistral-7B	🔓	<u>3.39</u>	3.81	**3.28**	2.98	**3.33**
Llama-3.1-8B	🔓	3.32	**3.89**	3.13	2.95	3.25
GPT-4o-mini	🔒	3.31	3.69	3.10	2.96	3.25
Gemini-2.0-flash	🔒	**3.42**	3.45	<u>3.15</u>	**3.02**	<u>3.32</u>
Claude-3.5-haiku	🔒	3.38	3.81	3.14	<u>2.99</u>	3.31

Table 3: LLM-as-a-Judge evaluation results using BM25 and Dense retrievers with TopK = 10 configuration. **Bold** and <u>underlined</u> values denote the best and second-best results for each metric.

Future Directions

1. **Temporary issues** (e.g., broken facilities, hygiene concerns, or construction noise) are often resolved.
   - Future works could integrate **temporal reasoning**.
2. **Incorporating star ratings**, lower-rated reviews often have negative aspects, can help in 'CONS' extraction.
3. **Promotional, spam, and manipulated reviews** reviewer's experience and credibility influence review quality.