Review of: "LFOSum: Summarizing Long-form Opinions with Large Language Models"

Dr. Ajay Rastogi[1]

[1] Lovely Professional University, Phagwāra, India

Potential competing interests: No potential competing interests to declare.

The paper presents *LFOSum*, a novel approach to summarizing long-form user reviews using Large Language Models (LLMs). It introduces a new dataset, scalable training-free summarisation methods, and novel reference-free evaluation metrics. The work addresses challenges of long-form summarisation, such as scalability, information overload, and the lack of annotated datasets. The proposed methods include the Long-form Critic and Retrieval-Augmented Generation (RAG) frameworks, alongside evaluation metrics that emphasize faithfulness, sentiment alignment, and opinion accuracy.

Strengths

Innovative Dataset:

- The LFOSum dataset, containing over 1,500 reviews per entity paired with critical summaries, is a significant contribution. Its focus on book-length inputs (>100,000 tokens) provides a new benchmark for long-form opinion summarisation.

Scalable Summarisation Techniques:

- The training-free approaches (Long-form Critic and RAG) enable scalable summarisation with control over sentiment, length, and queries. This adaptability adds practical value.

Novel Evaluation Metrics:

- Aspect-Opinion-Sentiment (AOS) triplet-based metrics offer fine-grained assessments of summary faithfulness, addressing a critical gap in the evaluation of sentiment-rich summaries.

Comprehensive Experimental Evaluation:

- The paper evaluates multiple LLMs (open and closed source) using diverse metrics, providing a holistic analysis of model performance under different configurations.

Focus on Faithfulness and Hallucination Prevention:

- The work emphasises reducing hallucinations in LLM outputs, an ongoing challenge in NLP.

Suggestions for Improvement

Expand Dataset Coverage:

- Include reviews in multiple languages to make the dataset and methods more inclusive and widely applicable.

Real-World Testing:

- Implement and analyze the methods in live settings, such as e-commerce or travel platforms, to validate performance and usability.

Enhance Open-Source Models:

- Investigate ways to improve the performance and reliability of open-source LLMs, such as better prompt engineering or fine-tuning strategies.

Incorporate Efficiency Metrics:

- Provide insights into computational costs, memory requirements, and runtime performance to assess scalability in practical deployments.

Focus on Diverse Retrieval Techniques:

- Explore alternative retrieval methods or hybrid approaches to further improve the RAG framework’s performance.

Broaden the Scope of Metrics:

- Extend evaluation metrics to include aspects like user satisfaction, coherence, and redundancy to capture more comprehensive performance indicators.

Overall, the paper makes a valuable contribution to the field of long-form opinion summarisation by addressing challenges of scalability and evaluation. Its novel dataset, methods, and metrics provide a foundation for future research. By addressing the identified weaknesses and implementing the suggested improvements, the impact and practical relevance of the work could be significantly enhanced.