Qeios

Peer Review

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

Arpita Gupta[1]

1. Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vijayawada, India

Strengths:

Relevance and Scope:

- The manuscript addresses a pressing problem in NLP: summarizing large-scale, long-form opinions effectively using advanced techniques such as RAG (Retrieval-Augmented Generation) and Long-Context LLMs.
- The introduction of a novel dataset, **LFOSum**, with over 1,000 reviews per entity paired with expert-annotated summaries, is a noteworthy contribution.

Innovative Contributions:

- Incorporation of **control mechanisms** (query, sentiment, length) for summarization allows customization and scalability.
- Development of novel **reference-free evaluation metrics** based on AOS (Aspect-Opinion-Sentiment) triplets provides a new approach to evaluate faithfulness in sentiment-heavy domains.

Comprehensive Evaluation:

- The benchmarking of multiple open-source and closed-source LLMs, along with a detailed analysis of challenges such as JSON parsing errors and sentiment adherence, is thorough and insightful.

Weaknesses:

Lack of Novelty in Methodology:

- While the dataset and evaluation framework are valuable, the summarization methods—RAG and Long-Context LLMs—rely heavily on established approaches without significant methodological innovation.
- The proposed methods are primarily application-focused rather than introducing fundamentally new techniques.

Validation Gaps:

doi.org/10.32388/456JFF

- The dataset's utility is not validated on real-world tasks beyond the presented experiments. For example, user studies or practical applications (e.g., e-commerce or travel platforms) are absent.
- The **AOS triplet-based evaluation metric**, though promising, lacks comparative validation against traditional metrics (e.g., ROUGE or BERTScore) in terms of reliability and scalability.

Data Limitations:

- The dataset focuses exclusively on the **hotel domain**, limiting generalizability to other areas such as e-commerce or entertainment.
- The reliance on TripAdvisor and Oyster reviews introduces potential biases due to platform-specific content styles and user demographics.

Evaluation Challenges:

- While the manuscript highlights the limitations of open-source LLMs, it does not adequately address the computational resource barriers of using closed-source, long-context LLMs like GPT-4o-mini and Claude-3-Haiku.
- Sentiment imbalance (e.g., scarcity of negative reviews) affects model evaluation, yet no mitigation strategies (e.g., data augmentation or rebalancing techniques) are proposed.

Presentation Issues:

- Several tables and figures lack proper explanations and readability:
  - **Table 2**: Scores for JSON adherence and parsing success are presented, but their implications for real-world use are unclear.
  - **Figures 3–5**: Prompt designs for the models are included but lack detailed discussion on their impact on results.
- Results on RAG verification metrics (e.g., Aspect Relevance, Sentiment Factuality, Opinion Faithfulness) are not contextualized with qualitative examples, making it harder to assess their practical relevance.

Writing Quality:

- The manuscript is overly verbose and includes redundant explanations (e.g., the dataset description is repeated multiple times).
- Grammatical errors and inconsistent terminology (e.g., switching between "length control" and "length adherence") affect readability.

Suggestions for Improvement:

Major Revisions:

Enhance Novelty:

- Propose methodological advancements, such as hybrid models combining RAG and fine-tuned LLMs or novel data augmentation strategies for long-form reviews.
- Explore innovative techniques for addressing challenges in JSON parsing and malformed outputs beyond the proposed Sketch-Fetch-Fill (SFF) approach.

Expand Validation:

- Validate the LFOSum dataset on additional domains (e.g., products, movies) to demonstrate its versatility.
- Include user studies to assess the practical utility of generated summaries in decision-making scenarios.

Strengthen Evaluation:

- Provide a comparative analysis of AOS triplet-based metrics with traditional metrics (e.g., ROUGE, BERTScore) to establish their reliability.
- Address sentiment imbalance through data augmentation or targeted loss functions in model training.

Improve Presentation:

- Clearly explain the significance of scores in **Tables 2 and 6**, particularly their implications for real-world applications.
- Include qualitative examples to illustrate the strengths and weaknesses of generated summaries, particularly in RAG verification.

Refine Writing:

- Streamline the manuscript to eliminate redundancy and improve clarity. Focus on concise explanations, particularly in the methodology and dataset sections.
- Address grammatical inconsistencies and ensure consistent terminology throughout the paper.

Conclusion:

The manuscript presents valuable contributions, particularly the LFOSum dataset and AOS triplet-based evaluation metrics. However, its reliance on established methods, limited validation, and presentation issues require significant revisions to meet the standards of high-impact research. By addressing these weaknesses, the authors can greatly enhance the paper's quality and relevance.

Declarations

Potential competing interests: No potential competing interests to declare.