VRTopic: Advancing Topic Modeling in Virtual Reality User Reviews with Large Language Models

1st Yijun Lu
Computer Science and Engineering
Waseda University
Tokyo, Japan
yijun@ruri.waseda.jp

2nd Haowei Cheng
Computer Science and Engineering
Waseda University
Tokyo, Japan
haowei.cheng@fuji.waseda.jp

3rd Jati H. Husen
Computer Science and Engineering
Waseda University
Tokyo, Japan
jati.h@asagi.waseda.jp

4th Hironori Washizaki
Computer Science and Engineering
Waseda University
Tokyo, Japan
washizaki@waseda.jp

5th Naoyasu Ubayashi
Computer Science and Engineering
Waseda University
Tokyo, Japan
ubayashi@aoni.waseda.jp

6th Nobukazu Yoshioka
Computer Science and Engineering
Waseda University
Tokyo, Japan
nobukazu@engineerable.ai

***Abstract*—With the rapid development of Virtual Reality (VR) technology, effectively understanding user feedback has become a core task for improving user experience and optimizing system functionality. However, extracting meaningful insights from VR user reviews remains challenging. Traditional topic modeling methods often generate unannotated and ambiguous topics, requiring extensive manual annotation and analysis. To address this issue, this study proposes an innovative approach that leverages state-of-the-art Large Language Models (LLMs) to automatically identify and precisely summarize key topics from VR user reviews. Ultimately, this research aims to generate accurate topics from VR-related textual inputs that genuinely reflect user concerns. By filling the gap in the application of LLMs to VR text analysis, this study provides VR developers with precise user insights, aiding product optimization and iterative improvement.**

I. Introduction

Topic modeling, a core technique in Natural Language Processing (NLP), has been widely applied across industries to analyze textual data. In software engineering, researchers extensively use methods such as LDA and BERTopic to analyze developer posts and user reviews, uncovering discussion content and trends. Our previous study on VR head-mounted displays (HMDs) development [1] applied BERTopic to cluster user discussions on VR HMDs, identifying key concerns. Notably, we found significant user worries regarding the safety of VR devices, providing valuable insights for VR developers. Sutton [2] highlighted how developers discuss the mechanisms behind technological success or failure, while Barua et al. [3] analyzed topic evolution over time. Similarly, past research has explored topic modeling in various software domains, such as machine learning [4], [5], mobile development [6], [7], security [8], and VR/video game development [9], [10].

While these studies highlight the effectiveness of traditional topic models in extracting meaningful insights, they face several challenges. 1) Traditional topic models often treat words as independent units, ignoring semantic relationships, which limits contextual understanding. 2) Additionally, the topics generated typically require significant manual annotation to become actionable, leading to high labor costs. 3) Furthermore, traditional models struggle to maintain performance when applied to large-scale or complex datasets, posing scalability issues.

The advent of Large Language Models (LLMs) has revolutionized topic extraction by addressing the limitations of traditional methods. Leveraging pre-trained embeddings and contextual understanding, LLMs generate semantically rich and actionable topics. For instance, Pham et al. [11] demonstrated that LLMs outperformed LDA in a Wikipedia-based study, achieving a harmonic mean purity of 0.74 compared to 0.64. Nori et al. [12] showed that LLMs achieved over 80% accuracy on BERTopic-annotated medical datasets. Zhang et al. [13] highlighted that LLMs generalized effectively across domains, surpassing traditional models in coherence and applicability. Similarly, Wang et al. [14] demonstrated LLMs' ability to summarize complex topics and identify emerging trends in scientific literature reviews. Finally, Wang et al. [15] used LLMs to extract topics from social media data, providing actionable insights for policymakers.

Despite these advancements, the use of LLMs in VR-related text analysis remains largely unexplored. Building on existing studies, this research proposes the development of an LLM-based topic extraction model specifically tailored to the VR domain. The model seeks to address the limitations of traditional methods by: 1) capturing deep semantic relationships in VR user feedback, 2) automating topic identification to reduce the need for manual intervention, and 3) providing actionable insights to developers for enhancing VR system development and optimizing user experiences. As shown in Figure 1.

II. Research Questions

This study focuses on addressing the following research questions (RQs) to explore the application of LLMs for extracting and analyzing VR-related topics.

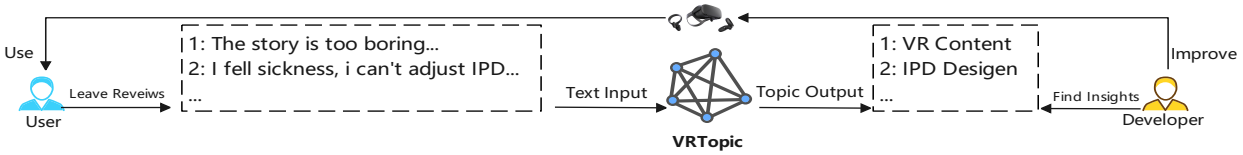

Fig. 1. Workflow

A. RQ1: How can LLMs be effectively applied to extract topics from VR-related textual data?

Motivation: Traditional topic models like LDA and BERTopic often fail to capture the nuanced semantics of VR feedback, including domain-specific jargon and complex expressions. LLMs offer a solution by leveraging advanced contextual understanding and language modeling.

Approach: Develop an LLM-based topic extraction framework tailored to the VR domain. Fine-tune pre-trained LLMs (e.g., GPT-4 [16], Flan [17], BLOOM [18], LLaMA [19]) on VR-specific datasets to enhance their understanding of domain-specific language. Employ prompt engineering and embeddings to improve topic granularity and relevance.

B. RQ2: How can the validity and reliability of the LLM-based topic extraction model be evaluated?

Motivation: Evaluating LLMs in domain-specific contexts like VR is critical due to unique challenges, such as complex terminology and diverse user feedback. Rigorous evaluation ensures the model's ability to extract accurate, actionable topics and identifies areas for improvement, enhancing its value for VR system development.

Approach: Building on the evaluation methods proposed by Kozlowskiet al. [13], this study employs two methods to assess the validity and accuracy of LLM-generated topics. 1. Matching with Predefined Labels. Domain experts manually create predefined labels as a baseline, and the LLM-generated topics are compared for overlap and semantic alignment. 2. Expert Scoring. Two authors independently score topics based on *Accuracy* (1–3, alignment with dataset context) and *Usefulness* (1–3, practical applicability). Inter-rater agreement is measured using Cohen's Kappa to ensure evaluation consistency.

C. RQ3: What are the key concerns and priorities of VR users as reflected in their feedback?

Motivation: Understanding user concerns and preferences is critical for optimizing VR systems and improving user experience. Identifying key topics in user feedback helps developers prioritize issues and design user-centric solutions.

Approach: Apply the LLM-based topic extraction framework to analyze user reviews collected from VR platforms (e.g., SteamVR, HTCVive). Extract and categorize topics into actionable domains such as technical challenges, user experience feedback, and feature requests.

D. RQ4: How do these topics and user concerns evolve over time?

Motivation: VR is a rapidly evolving field, and user expectations and concerns shift as technology advances. Analyzing temporal trends helps developers address persistent issues and anticipate future demands.

Approach: Building on the approach of Upp et al. [20], analyze both the absolute and relative influence of topics over time to identify significant changes and areas that require continued attention from developers across different time periods and VR platforms.

III. Results and Contributions

A. Expected Results

1) Cohen's Kappa value between evaluators is expected to exceed 0.8, confirming consistency and reliability.
2) LLM-generated topics are anticipated to achieve over 80% accuracy, surpassing traditional methods and human-defined baselines.
3) The model will effectively extract user concerns, summarizing them into actionable topics (e.g., "latency issues" or "privacy protection").
4) Temporal analysis will reveal evolving trends, such as declining focus on "motion sickness" and increasing discussions on "privacy," providing strategic guidance for developers.

B. Key Contributions

1) **LLM Framework for VR Feedback:** A model that accurately identifies VR user concerns (e.g., input: raw reviews, output: structured topics) to guide product optimization.
2) **Evaluation Framework:** Introduces metrics (e.g., accuracy, coherence) and qualitative methods (e.g., usefulness scoring, inter-rater consistency) for domain-specific LLM evaluation.
3) **Topic Evolution Insights:** Highlights user concern trends over time (e.g., reduced "motion sickness" concerns, increased focus on "privacy"), aiding resource prioritization.

IV. Question for DECS Committee

What additional metrics or validation techniques would you recommend to enhance the reliability of our LLM-based topic modeling approach?

How can we ensure fairness and minimize biases in LLM-generated topics for VR user feedback analysis?

Are there any opportunities for DECS participants to attend CHASE in person?

REFERENCES

[1] Y. Lu, K. Ota, and M. Dong, "An empirical study of vr head-mounted displays based on vr games reviews," *ACM Games*, vol. 2, no. 3, Aug. 2024. [Online]. Available: https://doi.org/10.1145/3665988

[2] M. Allamanis and C. Sutton, "Why, when, and what: analyzing stack overflow questions by topic, type, and code," in *2013 10th Working conference on mining software repositories (MSR)*. IEEE, 2013, pp. 53–56.

[3] A. Barua, S. W. Thomas, and A. E. Hassan, "What are developers talking about? an analysis of topics and trends in stack overflow," *Empirical software engineering*, vol. 19, pp. 619–654, 2014.

[4] A. A. Bangash, H. Sahar, S. Chowdhury, A. W. Wong, A. Hindle, and K. Ali, "What do developers know about machine learning: a study of ml discussions on stackoverflow," in *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*. IEEE, 2019, pp. 260–264.

[5] J. Han, E. Shihab, Z. Wan, S. Deng, and X. Xia, "What do programmers discuss about deep learning frameworks," *Empirical Software Engineering*, vol. 25, pp. 2694–2747, 2020.

[6] M. Linares-Vásquez, B. Dit, and D. Poshyvanyk, "An exploratory analysis of mobile development issues using stack overflow," in *2013 10th Working Conference on Mining Software Repositories (MSR)*. IEEE, 2013, pp. 93–96.

[7] C. Rosen and E. Shihab, "What are mobile developers asking about? a large scale study using stack overflow," *Empirical Software Engineering*, vol. 21, pp. 1192–1223, 2016.

[8] X.-L. Yang, D. Lo, X. Xia, Z.-Y. Wan, and J.-L. Sun, "What security questions do developers ask? a large-scale study of stack overflow posts," *Journal of Computer Science and Technology*, vol. 31, pp. 910–924, 2016.

[9] J. Dong, K. Ota, and M. Dong, "What are the points of concern for players about vr games: An empirical study based on user reviews in different languages," *ACM Games*, vol. 2, no. 4, Sep. 2024. [Online]. Available: https://doi.org/10.1145/3663739

[10] R. Epp, D. Lin, and C.-P. Bezemer, "An empirical study of trends of popular virtual reality games and their complaints," *IEEE Transactions on Games*, vol. 13, no. 3, pp. 275–286, 2021.

[11] C. M. Pham, A. Hoyle, S. Sun, P. Resnik, and M. Iyyer, "Topicgpt: A prompt-based topic modeling framework," *arXiv preprint arXiv:2311.01449*, 2023.

[12] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of gpt-4 on medical challenge problems," *arXiv preprint arXiv:2303.13375*, 2023.

[13] D. Kozlowski, C. Pradier, and P. Benz, "Generative ai for automatic topic labelling," *arXiv preprint arXiv:2408.07003*, 2024.

[14] D. Wang and S. Zhang, "Large language models in medical and healthcare fields: applications, advances, and challenges," *Artificial Intelligence Review*, vol. 57, no. 11, p. 299, 2024.

[15] Z. Wang, R. Li, B. Dong, J. Wang, X. Li, N. Liu, C. Mao, W. Zhang, L. Dong, J. Gao *et al.*, "Can llms like gpt-4 outperform traditional ai tools in dementia diagnosis? maybe, but not today," *arXiv preprint arXiv:2306.01499*, 2023.

[16] K. Sanderson, "Gpt-4 is here: what scientists think," *Nature*, vol. 615, no. 7954, p. 773, 2023.

[17] Z. Chen, K. Liu, Q. Wang, W. Zhang, J. Liu, D. Lin, K. Chen, and F. Zhao, "Agent-flan: Designing data and methods of effective agent tuning for large language models," *arXiv preprint arXiv:2403.12881*, 2024.

[18] H. Huang, Y. Feng, C. Shi, L. Xu, J. Yu, and S. Yang, "Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[19] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[20] G. Uddin, F. Sabir, Y.-G. Guéhéneuc, O. Alam, and F. Khomh, "An empirical study of iot topics in iot developer discussions on stack overflow," *Empirical Software Engineering*, vol. 26, pp. 1–45, 2021.