

INTISARI

Perkembangan *Large Language Models* (LLMs) telah membuka peluang baru dalam tugas *Natural Language Processing* (NLP), termasuk klasifikasi topik pada data media sosial yang bersifat tidak terstruktur dan sering mengandung lebih dari satu topik. Namun, efektivitas berbagai strategi *prompting* dalam meningkatkan performa klasifikasi masih perlu dievaluasi. Penelitian ini bertujuan membandingkan performa strategi *prompting* dan LLM dalam mengklasifikasikan komentar YouTube mengenai *residential solar photovoltaic* (PV). Penelitian menggunakan data berupa 10.192 komentar YouTube dari penelitian Widiyanto (2026). Sebanyak 40 komentar dipilih sebagai data evaluasi dan dianotasi oleh tujuh annotator berdasarkan 5 kategori topik dan 17 subtopik yang diadaptasi dari Shakeel et al. (2023). Reliabilitas anotasi dievaluasi menggunakan *Krippendorff's Alpha*, kemudian dilakukan adjudikasi untuk menghasilkan *validated label* sebagai *ground truth*. Tiga strategi *prompting*, yaitu *Basic Chain-of-Thought* (*Basic CoT*), *Minimal Chain-of-Thought* (*Minimal CoT*), dan *Heuristic-guided Chain-of-Thought* (*Heuristic-guided CoT*) diuji menggunakan GPT. Strategi *prompting* terbaik kemudian diterapkan pada GPT, Claude, dan LLaMA. Performa klasifikasi dievaluasi menggunakan *Exact Match Accuracy*, *Macro Precision*, *Macro Recall*, *Macro F1-score*, *Hamming Loss*, dan *Mean Jaccard*. Kombinasi terbaik selanjutnya diterapkan pada seluruh data untuk menganalisis distribusi topik, distribusi subtopik, keterkaitan antartopik, serta dibandingkan dengan hasil penelitian sebelumnya, yaitu Widiyanto (2026), yang menggunakan pendekatan BERTopic. Hasil penelitian menunjukkan bahwa strategi *Heuristic-guided CoT* memberikan performa terbaik. Pada pengujian antar-LLM, GPT menghasilkan performa terbaik berdasarkan nilai *Macro Recall* dan *Macro F1-score*. Kombinasi GPT dengan *Heuristic-guided CoT* ditetapkan sebagai kombinasi terbaik dan digunakan untuk mengklasifikasikan seluruh data. Hasil implementasi menunjukkan bahwa topik yang paling sering muncul dalam diskusi masyarakat mengenai *residential solar PV* adalah teknis, diikuti oleh ekonomi, regulasi, sosial, dan lingkungan. Analisis multilabel juga menunjukkan bahwa sebagian besar komentar membahas lebih dari satu topik secara bersamaan, dengan kombinasi teknis-ekonomi sebagai pasangan topik yang paling sering muncul. Berdasarkan perbandingan hasil penelitian ini dengan penelitian Widiyanto (2026) yang menggunakan BERTopic, diketahui bahwa kedua pendekatan mengidentifikasi aspek teknis sebagai topik yang paling banyak muncul dalam diskusi mengenai PLTS. Namun demikian, terdapat perbedaan persentase komentar pada topik ekonomi yang dipengaruhi oleh pendekatan klasifikasi yang digunakan.

Kata kunci: *Large Language Models*, *prompt engineering*, klasifikasi topik, *residential solar photovoltaic*, YouTube, *chain-of-thought*.

ABSTRACT

The development of Large Language Models (LLMs) has created new opportunities for various Natural Language Processing (NLP) tasks, including topic classification of social media text, which is often unstructured and may contain multiple topics within a single comment. However, the effectiveness of different prompting strategies in improving classification performance still requires further evaluation. This study aims to compare the performance of various prompting strategies and LLMs in classifying YouTube comments related to residential solar photovoltaic (PV) systems. The study utilized a dataset of 10,192 Indonesian YouTube comments obtained from Widiyanto (2026). A subset of 40 comments was selected as the evaluation dataset and annotated by seven annotators based on 5 topic categories and 17 subtopics adapted from Shakeel et al. (2023). Annotation reliability was assessed using Krippendorff's Alpha, followed by an adjudication process to produce validated labels as the ground truth. Three prompting strategies, namely Basic Chain-of-Thought (Basic CoT), Minimal Chain-of-Thought (Minimal CoT), and Heuristic-guided Chain-of-Thought (Heuristic-guided CoT), were evaluated using GPT. The best-performing prompting strategy was subsequently applied to GPT, Claude, and LLaMA. Classification performance was evaluated using Exact Match Accuracy, Macro Precision, Macro Recall, Macro F1-score, Hamming Loss, and Mean Jaccard. The best-performing combination was then applied to the entire dataset to analyze topic distribution, subtopic distribution, and inter-topic relationships, as well as to compare the results with a previous study, Widiyanto (2026), that employed the BERTopic approach. The results show that the Heuristic-guided CoT prompting strategy achieved the best overall performance. Among the evaluated LLM, GPT outperformed Claude and LLaMA based on Macro recall and Macro F1-score. Consequently, the combination of GPT and Heuristic-guided CoT was selected as the best-performing approach and applied to classify the entire dataset. The implementation results indicate that technical topics were the most frequently discussed aspects of residential solar PV, followed by economic, regulatory, social, and environmental topics. The multilabel analysis further revealed that most comments discussed more than one topic simultaneously, with the technical-economic combination being the most common. A comparison with the BERTopic-based study by Widiyanto (2026) showed that both approaches identified technical aspects as the dominant topic in public discussions on residential solar PV. Nevertheless, differences were observed in the proportion of comments classified into economic topics, reflecting the influence of the classification approach employed.

Keywords: Large Language Models, prompt engineering, topic classification, residential solar photovoltaic, YouTube, chain-of-thought.