Vol. 10, No. 1, April 2025, Page. 252-262
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

Does Personalization Matter in Prompting? A Case Study of Classifying Paper Metadata Using Zero-Shot Prompting

Chandra Lesmana[*1], Muhammad Okky Ibrohim[1,2] Indra Budi[1]

[1]Computer Science, Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia
[2]Institute of Social Studies, University of Tartu, Tartu, Estonia
Email: [1]chandra.lesmana21@ui.cs.ac.id, [1,2]okkyibrohim@cs.ui.ac.id, [1]indra@cs.ui.ac.id

Abstrak

Tinjauan pustaka sistematis atau *Systematic Literature Review* (SLR) merupakan salah satu cara peneliti untuk mendapatkan informasi perkembangan penelitian pada suatu topik secara testruktur. Hal tersebut menjadikan SLR sebagai metode yang disukai oleh para peneliti karena pada prosesnya melibatkan analisis dengan sistematis. Secara umum, ada tiga tahapan untuk melakukan SLR yaitu perencanaan, pelaksanaan, dan pelaporan. Namun dalam menyusun SLR membutuhkan waktu yang lama karena melewati seluruh tahapan satu per satu. Untuk mengatasi permasalahan tersebut, dibutuhkan proses otomasi sehingga dapat mempercepat proses penyusunan SLR. Penelitian sebelumnya telah melakukan proses otomasi berupa klasifikasi dokumen SLR dengan memanfaatkan beberapa model pembelajaran mesin yang membutuhkan banyak data latih seperti *Naïve Bayes*, *Support Vector Machine*, dan *Logistic Model Tree*. Pada penelitian ini, penulis melakukan proses otomasi dengan memanfaatkan *open-source Large Language Model* (LLM) yaitu Mistral-7B-Instruct-v0.2 dan LLaMA-3.1–8B-Instruct untuk melakukan klasifikasi judul dan abstrak dokumen SLR. Kami membandingkan pengaruh pemberian personalisasi pada *zero-shot prompting*. Dengan menggunakan LLM menggunakan *zero-shot prompting*, proses klasifikasi tidak lagi membutuhkan data latih sehingga tidak membutuhkan biaya anotasi data. Hasil eksperimen menunjukkan pemberian personalisasi meningkatkan performa klasifikasi, mendapatkan hasil terbaik dengan Macro F_1 0.5538 dengan menggunakan model Llama 3.1.

Kata kunci: tinjauan pustaka sistematis, klasifikasi metadata dokumen, model bahasa besar, *zero-shot prompting*, personalisasi

Abstract

Systematic Literature Review (SLR) is one way for researchers to obtain information on research developments on a topic in a structured manner. This makes SLR a preferred method by researchers because the process involves systematic, objective analysis and focuses on answering research questions. In general, there are three stages to conducting SLR, namely planning, implementation, and reporting. However, compiling an SLR takes a long time because it goes through all the stages one by one. To overcome this problem, an automation process is needed so that it can speed up the SLR compilation process. Previous studies have carried out an automation process in the form of SLR document classification by utilizing several machine learning models that require a lot of training data like Naïve Bayes, Support Vector Machine, and Logistic Model Tree. In this study, the authors conducted an automation process by utilizing open-source Large Language Model (LLM) namely Mistral-7B-Instruct-v0.2 and LLaMA-3.1–8B to classify title and abstract of SLR documents. We compared the effect of using personalization on zero-shot prompting. By using LLM with zero-shot prompting, the classification process no longer requires training data, so that it does not need data annotation cost. Experiment results showed that personalization improved classification performance, getting the best results with Macro F_1 0.5538 using the Llama 3.1 model.

Keywords: *systematic literature review, paper metadata classification, large language model, zero-shot prompting, personalization*

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

1. INTRODUCTION

A systematic literature review (SLR) is a method that researchers employ to develop a comprehensive understanding of the evolution of a research topic in a systematic, objective, and evidence-based manner. [1]. In the context of its implementation, SLR is defined by three distinct phases: planning, conducting, and reporting [2]. The conducting stage is of particular importance because it is during this stage that the document selection process is carried out based on specific criteria, such as the relevance of the content to the research question. One big challenge in getting SLR is selecting the right documents. This process is repetitive and time-consuming, especially when there are a lot of documents to be analyzed [3]. Therefore, it is important to automate document selection to improve efficiency and reduce the workload of researchers.

Previous research has attempted to automate document selection using classic machine learning models, such as Naive Bayes (NB) [4], Support Vector Machine (SVM) [4], [5], and Logistic Model Tree (LMT) [4]. These classic machine learning methods have shown good results, but they need model training process with data. Therefore, this method is not ideal for automating the selection of SLR documents because each SLR contains different topics, requiring the data to be re-annotated for each new SLR process. To solve this problem, we need a method that can automate the selection of SLR documents without going through a training process, one of those things is using a Large Language Model (LLM).

Large Language Model (LLM) is a model created using large, complex data sets that enables it to recognize and understand text naturally [6]. LLM outperform traditional machine learning methods because they are trained on larger data sets and more complex architectures [7]. The in-context learning (ICL) paradigm enables us to perform classification with an LLM without modifying its weights. [8]. In ICL, zero-shot prompting can perform classification without providing the model with examples, only instructions [9].

The zero-shot prompting approach allows LLM to perform tasks without requiring additional training on specific data, provided they are given clear instructions in the form of prompts. Consequently, prompt engineering emerges as a pivotal strategy for improving LLM performance and facilitating the automation of document selection processes in SLR. Several earlier studies have demonstrated the effectiveness of LLM in this context [10]. Dennstädt et al. [11] in 2024 was conducted zero-shot prompts to automate SLR document classification. However, the prompts used seem simple. This means that when carrying out the task, the model could possibly not limit its knowledge. This could result in the model not focusing on the domain and topic of the SLR to be classified.

The personalization prompt is an attempt to modify the instructions to give the model a viewpoint, hopefully producing a LLM response suitable for the user needs [12]. The method is to provide a role such as "You are a research assistant..." in the prompt so that the model can presume he is assuming the given role. According to several studies, adding personalization to prompts can improve the models performance [13], [14].

To overcome these problems, we study to examines the performance of SLR document classification using LLM by applying personalization to the prompt. The personalization gives directed into prompt to act as a domain researcher conducting the SLR. The purpose of providing personalization is that the LLM is expected to be able to limit its knowledge according to the given persona so that it can perform its tasks more effectively. To test this approach, the study used a dataset in the form of a collection of metadata in the form of titles and abstracts that had been manually given selection labels ('relevant' and 'irrelevant') by by Ibrohim et al,[1] [15] in natural environment sentiment analysis topic.

This article explains how the research was done and what the results were. The introduction explains the background of the problem and why it is important to study it, along with a review of previous studies that relate to the current study. The research method is explained in detail, including the approach, data collection techniques, and analysis procedures used. The last part of the article shows the results of the experiment and talks about them. It also connects these results to the original goals of the research and previous findings. This helps show what the results mean and how they add to what we already know.

[1] https://github.com/okkyibrohim/slr-sa-in-ne/tree/main

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

2. RELATED WORKS

Several studies have shown that machine learning algorithms, such as Support Vector Machines (SVM), Naïve Bayes (NB), and Logistic Regression (LR), can greatly reduce the amount of work required for manual screening. For example, Kebede et al. [16] reported a 36% false positive rate, indicating a trade-off between sensitivity and specificity. In a similar case, Bao et al. [17] conducted a comparative analysis of Support Vector Machine (SVM) and Convolutional Neural Network (CNN) models in the context of biomedical abstract classification. The study observed comparable accuracies of approximately 89% across both models. However, it was noted that training both models required a lot of labelled data for them to perform well. A key problem with these methods is that they require a lot of work to create and a lot of high-quality, labelled data, which can be very time-consuming.

Previous research by Clavié et al. [13] examined the application of prompt engineering techniques for job classification with LLM, particularly GPT-3.5-turbo. One tested approach was adding role-personalization elements to the prompt, such as instructing the model to "act as a professional recruiter". The results showed that using personalization can improve the model performance compared to using zero-shot without personalization.

Recently, Dennstädt et al. [11] conducted an exploratory study using various publicly available LLM, including FlanT5, OpenHermes, Mixtral, and Platypus2, to evaluate the relevance of biomedical publications based on their titles and abstracts. The study used structured prompts and Likert-scale scoring. This approach yielded high sensitivity but variable specificity across different datasets. Thus, it highlighted the potential and limitations of LLM in screening tasks. They study did not incorporate personalization techniques. The absence of personalization suggests an area for future research.

Table 1 Summary of related work

Author	Research Method	Model
[16]	Comparing to classify title and abstract of SLR document	Naïve Bayes, Support VectorMachines (SVM), regularized logistic regressions, neural networks, random forest,Logit boost, and XGBoos
[17]	Comparing analysis 2 models for title and abstract classification	SVM, CNN
[11]	Classify title and abstract of biomedical SLR document	FlanT5, OpenHermes, Mixtral, Platypus2
[13]	Comparing 2 LLM performance to classify title and abstract of SLR document	GPT-3.5-turbo, Llama-70B

Based on Table 1, previous studies on automating document selection in SLR have predominantly relied on traditional machine learning (ML) and deep learning (DL) models. While these approaches have shown reasonable performance, they typically require extensive feature engineering, manual annotation, and domain-specific preprocessing, which limit their scalability and adaptability across different research domains. More recent developments have begun exploring the use of LLM for document classification in SLR due to their superior language understanding and generalization capabilities. However, they have relied on relatively simple prompt designs, often lacking contextual depth or domain specificity. This study addresses these gaps by introducing a personalized prompting strategy, in which the LLM is assigned a specific role such as a "research assistant" to enhance its contextual comprehension and alignment with the task. By integrating personalization into the prompt, the proposed approach aims to improve classification accuracy and better simulate domain expertise.

3. RESEARCH METHOD

In general, our research flow conducted in Figure 1.

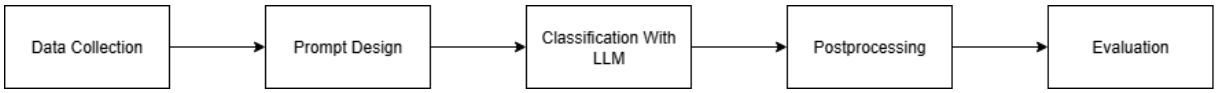

Figure 1. Research flow

3.1 Data Collection

The first step of this study involves the collection and utilization of a test dataset. The dataset employed was sourced from a SLR conducted by Ibrohim et al., which specifically focused on sentiment analysis in the context of environmental issues [15]. The selection of this dataset was made based on two key factors. First, there was thematic alignment with the scope and objectives of the study. Second, the dataset demonstrated a clear and transparent annotation methodology, which is a crucial element in the analysis and interpretation of data. Each document in the dataset includes title and abstract which provide sufficient textual context for relevance classification tasks. These documents are further labeled with binary classifications (relevant or irrelevant), offering a clear ground truth for evaluating the performance of automated classification models.

The dataset comprises a total of 1435 documents, with 90 labelled as relevant and the remaining 1345 categorized as irrelevant. This significant imbalance in class distribution realistically reflects the challenges typically encountered in systematic review tasks, where relevant studies often constitute a small fraction of the total literature. Addressing this imbalance is critical in evaluating the robustness of classification models, especially in domains with high information redundancy and noise. Moreover, the public availability of the dataset through a GitHub repository enhances its transparency and reproducibility, allowing other researchers to replicate or extend the findings of this study.

3.2 Prompt Design

This study incorporates personalization into the prompt to improve the contextual comprehension of LLM. Personalization, in this context, refers to the deliberate assignment of a specific persona or role to the model to influence its behavior and responses. For instance, by assigning the role of a "research assistant," the model is encouraged to interpret queries and generate responses with the tone, logic, and structure expected from someone with academic or domain-relevant expertise. This technique enables the model to better understand the user's intent and respond more accurately, especially when dealing with tasks that require domain-specific reasoning. The core idea is to leverage the model's ability to simulate domain expertise through role-based cues in the prompt, thereby improving accuracy, maintaining consistency, and reducing the likelihood of ambiguous or generic outputs [18]. In complex research workflows, this kind of personalization can help guide the model toward more purposeful and context-aware decision-making, particularly in tasks like classification.

In line with this, the research employs a zero-shot prompting approach to automate document classification during the systematic literature review (SLR) phase. This method utilizes the LLM's general language understanding capabilities without requiring additional fine-tuning or manually annotated training datasets, thereby simplifying the deployment process and enhancing adaptability across topics [9]. As illustrated in Figure 2, the prompt is designed to include a concise task description, clearly stated relevance and irrelevance criteria, and the input metadata in the form of document titles and abstracts. This structured approach aims to provide sufficient context for the model to make informed binary classification decisions. By integrating personalization and zero-shot prompting, the study demonstrates a scalable and efficient method for leveraging LLMs in automating labour-intensive processes within evidence-based research synthesis.

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

Task:
You are a researcher assistant in the sentiment analysis for natural environment field with experience screening titles and abstracts in Systematic Literature Review (SLR). Determine whether this scientific paper is 'relevant' or 'irrelevant' according to the relevant/irrelevant criteria from the provided input paper metadata.

Instructions:
- Please only answer 'relevant' or 'irrelevant' without any explanation.

Relevant/Irrelevant Criteria:
- criteria_1
  - relevant_criteria
  - irrelevant_criteira
- criteria_2
  - relevant_criteria
  - irrelevant_criteira
- criteria_3
  - relevant_criteria
  - irrelevant_criteira

Input Paper Metadata:
- Title: {title}
- Abstract: {abstract}

Answer:

Figure 2. Prompt used to classify the metadata. See Table 2 in [15] for details of the relevant/irrelevant criteria.

3.3 Classification With LLM

To execute the prompt, this study utilizes open-source LLM to facilitate the automated classification of scientific documents. The classification task is conducted using two decoder-only models that have undergone instruction tuning: LLaMA-3.1–8B-Instruct[2] [19] and Mistral-7B-Instruct-v0.2[3] [20]. These models were selected based on a combination of criteria, including their proven performance in previous evaluations [21]. Instruction tuning plays a vital role in enhancing the models' responsiveness to specific tasks by training them to follow natural language instructions more effectively. This characteristic aligns well with the zero-shot prompting approach used in this study, where models must interpret task instructions directly from the prompt without additional fine-tuning.

The models were deployed locally on a high-performance machine equipped with an NVIDIA A100 GPU, enabling efficient parallel computation and large-scale inference. By implementing the LLM in a local environment rather than relying on cloud-based APIs, the study gains full control over the inference pipeline, allowing for more consistent experimentation, reduced operational latency, and enhanced data privacy. Once the structured prompt is generated, it is passed to the LLM, which interprets the instructions, applies the relevance criteria, and returns a classification output.

3.4 Postprocessing

After obtaining the classification response from the LLM, a dedicated postprocessing stage is applied to extract and standardize the output. This stage plays a crucial role in ensuring the reliability of the automated classification pipeline. Although the prompt is carefully crafted to instruct the model to respond strictly with either "relevant" or "irrelevant," LLM may occasionally produce outputs that include additional tokens, formatting variations, or language artifacts. Therefore, postprocessing is necessary to enforce strict conformity to the expected output format. This process guarantees that only valid, interpretable labels are retained, which is essential for maintaining the integrity of downstream evaluation metrics.

To operationalize this, a regular expression (RegEx) filter is employed to parse the model outputs. The RegEx is designed to match only the two allowable classification terms, ignoring any extraneous

[2] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[3] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

content that might inadvertently be included in the response. This filtering step minimizes the risk of misclassification caused by inconsistent formatting or ambiguous wording. Moreover, it standardizes the outputs into a clean binary form, which is necessary for accurate performance evaluation. By validating and sanitizing the LLM-generated labels, the postprocessing stage strengthens the overall reliability and robustness of the classification workflow.

Table 2 Example of postprocessing

Paper Metadata	Response	Filtered Response	Final Response
[24]	`Relevant Justification: - The paper discusses environmental topics (online-reviews domain)`	Relevant	relevant
[27]	`irrelevant Justification: The paper does not discuss sentiment analysis modeling, nor does it propose an N`	irrelevant	irrelevant

3.5 Evaluation

The evaluation strategy of this study is predicated on the use of Macro F_1 as the principal metric to assess the classification performance of LLM in the context of document selection for SLR. Macro F_1 was deliberately chosen because it provides a balanced evaluation of performance across classes, regardless of their frequency in the dataset. This is particularly important in SLR scenarios, where relevant documents often represent only a small minority of the total corpus, resulting in a pronounced class imbalance. Unlike micro-averaged metrics, which may be disproportionately influenced by the majority class, the Macro F_1 computes the F_1 for each class independently and then averages the results. This ensures that both the "relevant" and "irrelevant" classes are given equal weight in the final performance evaluation, providing a more nuanced understanding of how well the model performs across the classification spectrum[22].

To assess the robustness and statistical significance of the observed classification outcomes further, McNemar test is used for hypothesis testing. This non-parametric method is specifically designed to evaluate paired nominal data, making it suitable for comparing the predictions of two models on the same dataset [23]. In this study, McNemar's test is used to determine the difference in performance when personalization is used versus when it is not. By focusing on instances where the models disagree, McNemar's test reveals whether improvements in MacroF_1 are due to genuine performance gains. This statistical validation step strengthens the credibility of the study's conclusions and adds rigor to the experimental analysis.

4. RESULTS AND DISCUSSION

Experiments were conducted by testing four scenario combinations varying two main parameters, namely (1) the use of role personalization in the prompt (yes or no), and (2) the type of LLM model used (LLaMA or Mistral). Performance evaluation was conducted using the Macro F_1 metric, as this metric provides fair assessment under unbalanced class conditions. Table 1 shows the experimental scores for each scenario.

Table 3 Each approach experiment result

Model	Personalization	Relevant			Irrelevant			*Macro F_1*
		Precision	Recall	F_1	Precision	Recall	F_1	
Llama-3.1-8B-Instruct	No	0.1158	0.6444	0.1963	0.9657	0.6706	0.7916	0.4939
	Yes	0.1628	0.7778	0.2692	0.9801	0.7323	0.8383	**0.5538**
Mistral-7B-Instruct-v0.2	No	0.7140	1.0000	0.1333	1.0000	0.1301	0.2303	0.1818
	Yes	0.8200	0.9667	0.1512	0.9920	0.2758	0.4316	0.2914

From the Table 1 show results of the experiment results demonstrate that give personalization into prompt can enhances the classification performance of both evaluated language models, though the degree of improvement varies. Llama-3.1-8B-Instruct demonstrate higher Macro F_1 which gets a Macro F_1 score 0.5538. This means that the model responds well to personalization. Mistral-7B-Instruct-v0.2 model shows lower than Llama-3.1-8B-Instruct even after personalization is added. This suggests that personalization improves the model ability to understand its task. On the other hand, when there was no personalization, performance suffered. This is probably because the model cannot understand the instructions, which leads to confusion.

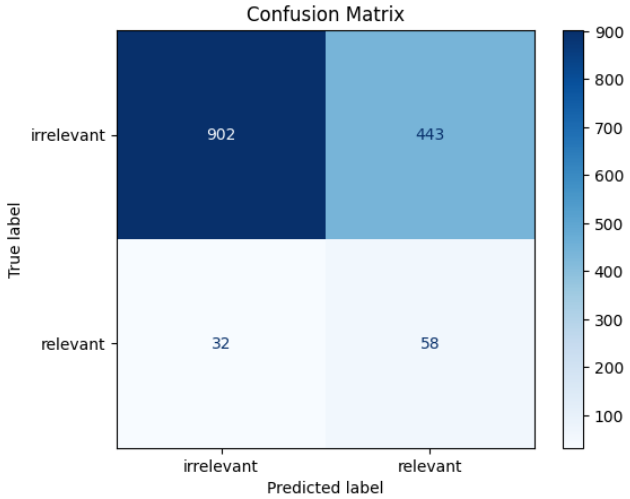

Figure 3. Llama-3.1-8B-Instruct without personalization approaches confusion matrix

The confusion matrix in Figure 3 shows how Llama 3.1 without personalization can tell the difference between relevant and irrelevant documents. The model correctly identified 902 irrelevant documents, but also incorrectly identified 443 documents as relevant (false positive). For the relevant class, the model correctly recognized 58 documents but failed to detect 32 others (false negatives). This shows that while the model is quite good at filtering out irrelevant documents, it still has trouble detecting relevant documents. This can lead to the loss of important literature during the initial selection process.

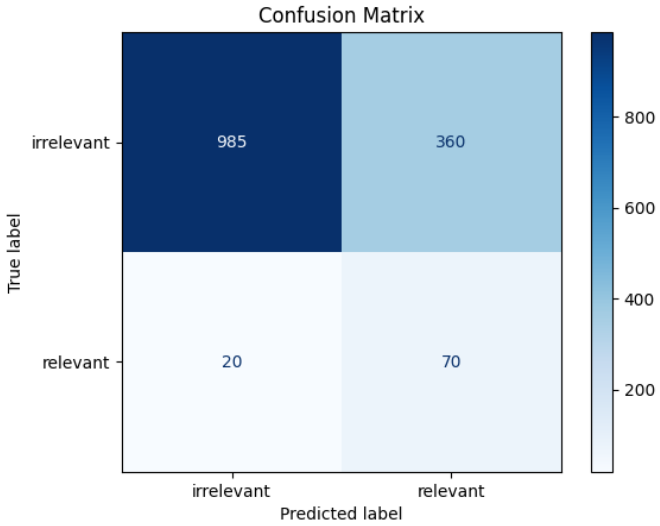

Figure 4. Llama-3.1-8B-Instruct with personalization approaches confusion matrix

The application of personalization to the prompt produces variations in the confusion matrix results, as illustrated in Figure 4. Compared with Figure 3, there are changes that indicate improvements in classification prediction. With this approach, model correctly 985 irrelevant documents and 70 relevant

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

documents. There were 83 irrelevant documents and 12 relevant documents that were successfully corrected when implementing personalization. To evaluate the significance of personalization, a McNemar test was employed. The result of this analysis indicates that the application of personalization achieves a p-value of 1.034×10^{-14}, where the result is statistically significant at the 0.05 level of significance.

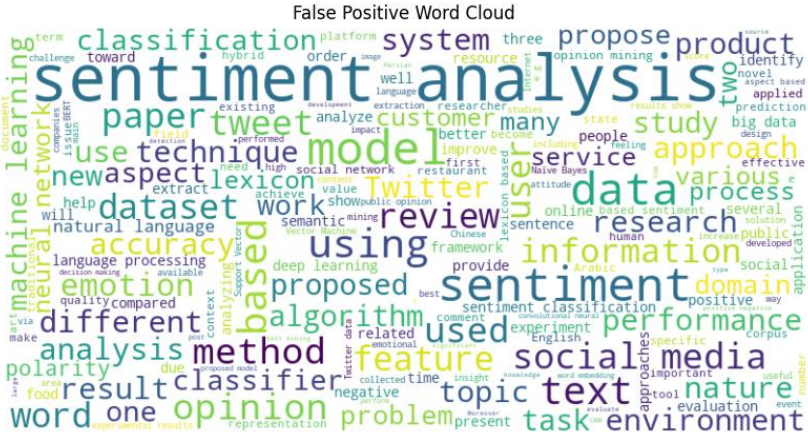

Figure 5. Word cloud from false positive

From Figure 4 show the term of "sentiment analysis" appears frequently in the "False Positive" word cloud. This indicates that the model often incorrectly identifies documents containing this term as relevant, even though they are irrelevant. This shows that the model depends too much on popular technical keywords without understanding how they are used. The term "sentiment analysis" is often used in documents that focus on technical aspects of natural language processing (NLP), like developing algorithms or evaluating methods. It is not often used to address big issues like public opinion on policies or social issues. This means that the model often fails to tell the difference between using terms as part of a technical study and using them for thematic analysis. This error indicates a fundamental limitation in the model's capacity to comprehend the context in great specificity. It needs a better way to understand the meaning of words. This will help it recognize if tools like sentiment analysis are used in a way that is related to the topic of study or if they are only mentioned in a general way.

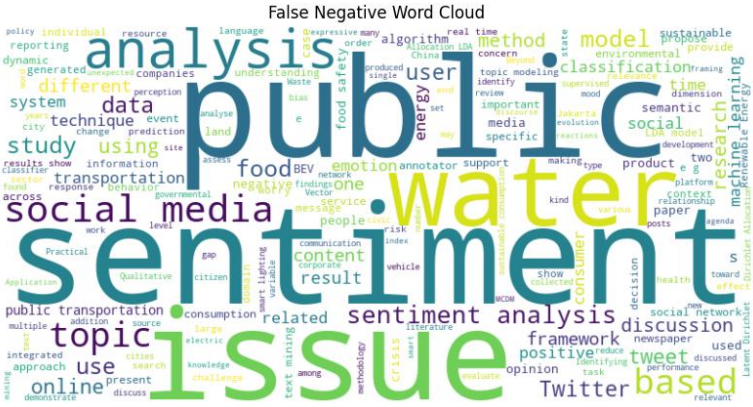

Figure 6. Word cloud from false negative

The word cloud in the "False Negative" category from Figure 5 shows that terms such as "public", "sentiment", "issue", and "water" are very dominant. This means that documents in this category contain information about social and environmental issues, but the model do not recognize them as relevant. The words "public" and "issue" are used a lot in the study because it focuses on public opinion, policy issues, and certain social problems. These should be included in the relevant classification. However, the content might be considered irrelevant because it lacks technical terminology or explicit methods, as seen in the model training documents. This error shows that the model does not understand situations where a topic is clearly explained but not always use special words that show it is relevant. Table 2 show

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

some examples of classification error metadata and LLM response from "False Positive" and "False Negative".

Tabel 1 Example of classification error. We do not provide the title and abstract details to avoid plagiarism. See the reference in the table to read the title and abstract.

Paper Metadata	Response	Error Type
[24]	Relevant Justification: - The paper discusses environmental topics (online-reviews domain)	False Positive
[25]	Relevant Justification: The paper discusses environmental topics (social microblogging sites, user views	False Positive
[26]	irrelevant. - Title: Land-water-food nexus of biofuels: Discourse and policy debates	False Negative
[27]	irrelevant Justification: The paper does not discuss sentiment analysis modeling, nor does it propose an N	False Negative

The model does not show any signs of hallucinations in the responses it produces. Each answer is always labelled as relevant or irrelevant, and an explanation is provided to support the label. So, the responses that the model produces are not only label-appropriate but also explain the answer. This shows that the model output is not filled with hallucinations.

5. CONCLUSION

Research has examined the Large Language Models (LLM) to automate of document selection in a Systematic Literature Review (SLR), especially the document classification stage based on title and abstract. The zero-shot prompting technique was chosen as the main method to utilize the capabilities of LLM without requiring training data or special annotations. A series of experiments were conducted by testing two popular open-source models, namely Meta-LLaMA-3.1-8B-Instruct and Mistral 7B-Instruct, with adding the personalization prompts and performance evaluation using the Macro F_1 metric on a dataset of 1435 documents with unbalanced class distribution. The findings indicated add personalization into prompt shown optimal performance, attaining a Macro F_1 score 0.5538 with Llama 3.1 8B model. This methodological approach can improve the model comprehension of instructions. The proposed approach has the potential to expedite the literature review process while maintaining reasonable performance, eliminating the need for retraining or additional annotation data.

Future research should investigate a wider range of prompt engineering techniques to improve the flexibility and effectiveness of language models. One such technique is few-shot prompting, which provides minimal examples to guide model behaviour. Exploring these approaches could reduce the need for extensive fine-tuning while improving task-specific performance. Furthermore, testing these methods on diverse datasets is crucial for assessing their robustness and improving the model generalizability in various real-world scenarios. To enable the model to understand the context in more depth, one strategy that can be applied is to use LLM with larger parameter sizes. Large-sized models usually perform better when handling complicated instructions and comprehending detailed situations. However, using this kind of model requires enough memory on the GPU. When there are hardware limitations, one option is to use models that have been trained with relevant information for the research topic or domain. These specialized models are good at identifying unique language terms and patterns. This makes them better at classifying things, even if they are smaller. To implement this approach for

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

application, future research can develop a general SLR document classification system that can be used by researchers from various disciplines, including but not limited to the field of computer science. This objective is to promote a more inclusive, efficient, and accessible SLR process across various fields.

ACKNOWLEDGEMENT

The author is very thankful to the UI-Tokopedia AI Center for the help with the computing facilities that were provided during this research process. The DGX-A100 device has accelerated the process of running the experiment. This helps make our research run smoothly.

REFERENCE

[1] G. Lame, "Systematic Literature Reviews: An Introduction," *Proc. Des. Soc. Int. Conf. Eng. Des.*, vol. 1, no. 1, pp. 1633–1642, Jul. 2019, doi: 10.1017/dsi.2019.169.
[2] Y. Xiao and M. Watson, "Guidance on Conducting a Systematic Literature Review," *J. Plan. Educ. Res.*, vol. 39, no. 1, pp. 93–112, Mar. 2019, doi: 10.1177/0739456X17723971.
[3] A. Chapman *et al.*, "Overcoming challenges in conducting systematic reviews in implementation science: a methods commentary," *Syst. Rev.*, vol. 12, no. 1, p. 116, Jul. 2023, doi: 10.1186/s13643-023-02285-3.
[4] H. Almeida, M.-J. Meurs, L. Kosseim, and A. Tsang, "Data Sampling and Supervised Learning for HIV Literature Screening," *IEEE Trans. NanoBioscience*, vol. 15, no. 4, pp. 354–361, Jun. 2016, doi: 10.1109/TNB.2016.2565481.
[5] A. Bannach-Brown *et al.*, "Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error," *Syst. Rev.*, vol. 8, no. 1, p. 23, Dec. 2019, doi: 10.1186/s13643-019-0942-7.
[6] H. Naveed *et al.*, "A Comprehensive Overview of Large Language Models," Nov. 23, 2023, *arXiv*: arXiv:2307.06435. Accessed: Dec. 25, 2023. [Online]. Available: http://arxiv.org/abs/2307.06435
[7] Y. Chang *et al.*, "A Survey on Evaluation of Large Language Models," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 3, pp. 1–45, Jun. 2024, doi: 10.1145/3641289.
[8] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," Jul. 22, 2020, *arXiv*: arXiv:2005.14165. Accessed: May 27, 2024. [Online]. Available: http://arxiv.org/abs/2005.14165
[9] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, in NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2022.
[10] X. Luo *et al.*, "Potential Roles of Large Language Models in the Production of Systematic Reviews and Meta-Analyses," *J. Med. Internet Res.*, vol. 26, p. e56780, Jun. 2024, doi: 10.2196/56780.
[11] F. Dennstädt, J. Zink, P. M. Putora, J. Hastings, and N. Cihoric, "Title and abstract screening for literature reviews using large language models: an exploratory study in the biomedical domain," *Syst. Rev.*, vol. 13, no. 1, p. 158, Jun. 2024, doi: 10.1186/s13643-024-02575-4.
[12] J. Chen *et al.*, "When large language models meet personalization: perspectives of challenges and opportunities," *World Wide Web*, vol. 27, no. 4, p. 42, Jul. 2024, doi: 10.1007/s11280-024-01276-1.
[13] B. Clavié, A. Ciceu, F. Naylor, G. Soulié, and T. Brightwell, "Large Language Models in the Workplace: A Case Study on Prompt Engineering for Job Type Classification," in *Natural Language Processing and Information Systems*, E. Métais, F. Meziane, V. Sugumaran, W. Manning, and S. Reiff-Marganiec, Eds., Cham: Springer Nature Switzerland, 2023, pp. 3–17.
[14] A. Salemi, S. Mysore, M. Bendersky, and H. Zamani, "LaMP: When Large Language Models Meet Personalization," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 7370–7392. doi: 10.18653/v1/2024.acl-long.399.
[15] M. O. Ibrohim, C. Bosco, and V. Basile, "Sentiment Analysis for the Natural Environment: A Systematic Review," *ACM Comput. Surv.*, vol. 56, no. 4, pp. 1–37, Apr. 2024, doi: 10.1145/3604605.

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

[16] M. M. Kebede, C. Le Cornet, and R. T. Fortner, "In-depth evaluation of machine learning methods for semi-automating article screening in a systematic review of mechanistic literature," *Res. Synth. Methods*, vol. 14, no. 2, pp. 156–172, 2023, doi: https://doi.org/10.1002/jrsm.1589.
[17] Y. Bao *et al.*, "Using Machine Learning and Natural Language Processing to Review and Classify the Medical Literature on Cancer Susceptibility Genes," *JCO Clin. Cancer Inform.*, no. 3, pp. 1–9, Dec. 2019, doi: 10.1200/CCI.19.00042.
[18] A. Salemi, S. Mysore, M. Bendersky, and H. Zamani, "LaMP: When Large Language Models Meet Personalization," Jun. 05, 2024, *arXiv*: arXiv:2304.11406. doi: 10.48550/arXiv.2304.11406.
[19] A. Dubey *et al.*, "The Llama 3 Herd of Models," Aug. 15, 2024, *arXiv*: arXiv:2407.21783. Accessed: Sep. 11, 2024. [Online]. Available: http://arxiv.org/abs/2407.21783
[20] A. Q. Jiang *et al.*, "Mistral 7B," Oct. 10, 2023, *arXiv*: arXiv:2310.06825. Accessed: Sep. 04, 2024. [Online]. Available: http://arxiv.org/abs/2310.06825
[21] A. López-Pineda, R. Nouni-García, Á. Carbonell-Soliva, V. F. Gil-Guillén, C. Carratalá-Munuera, and F. Borrás, "Validation of large language models (Llama 3 and ChatGPT-4o mini) for title and abstract screening in biomedical systematic reviews," *Res. Synth. Methods*, vol. 16, no. 4, pp. 620–630, Jul. 2025, doi: 10.1017/rsm.2025.15.
[22] K. Takahashi, K. Yamamoto, A. Kuchiba, and T. Koyama, "Confidence interval for micro-averaged F (1) and macro-averaged F (1) scores.," *Appl. Intell. Dordr. Neth.*, vol. 52, no. 5, pp. 4961–4972, Mar. 2022, doi: 10.1007/s10489-021-02635-5.
[23] M. Q. R. Pembury Smith and G. D. Ruxton, "Effective use of the McNemar test," *Behav. Ecol. Sociobiol.*, vol. 74, no. 11, Nov. 2020, doi: 10.1007/s00265-020-02916-y.
[24] Z. Gerolemou and J. Scholtes, "Target-Based Sentiment Analysis as a Sequence-Tagging Task," Nov. 2019.
[25] B. Khanna N, S. Moses J, and N. M, "SoftMax based User Attitude Detection Algorithm for Sentimental Analysis," *Procedia Comput. Sci.*, vol. 125, pp. 313–320, 2018, doi: 10.1016/j.procs.2017.12.042.
[26] L. L. Benites-Lazaro, L. L. Giatti, W. C. Sousa Junior, and A. Giarolla, "Land-water-food nexus of biofuels: Discourse and policy debates in Brazil," *Environ. Dev.*, vol. 33, p. 100491, Mar. 2020, doi: 10.1016/j.envdev.2019.100491.
[27] M. Altaweel and C. Bone, "Applying content analysis for investigating the reporting of water issues," *Comput. Environ. Urban Syst.*, vol. 36, no. 6, pp. 599–613, Nov. 2012, doi: 10.1016/j.compenvurbsys.2012.03.004.