Enhancing Content Categorization Systems: A Comparative Study of Integrated Diverse Datasets for Improved Adaptability

Rithika Florian Johnson
University of Massachusetts Amherst
rflorianjohn@umass.edu

ABSTRACT

This research introduces a novel methodology for achieving personalized news categorization by leveraging the LaMP architecture. The approach involves meticulous data preprocessing, utilizing a new personalization dataset, and incorporating existing datasets such as AG News and Book Depository. The integration fosters a comprehensive dataset, capturing diverse user interests and enhancing the model's adaptability to a wide range of content domains.

The core of the methodology lies in the integration of diverse datasets, which provides a holistic view of user preferences. The combination of the LaMP architecture with the BM25 algorithm for document retrieval ensures that the personalized prompts constructed are rich and reflective of nuanced user preferences. This strategic approach sets the stage for accurate predictions of content categories by the language model, specifically the Flan-T5-base-finetuned model.

The finetuning process using the generated prompts after retrieval enhances the language model's ability to discern and predict content categories accurately. Through a comparative analysis under various integration scenarios, the research sheds light on the advancements facilitated by the LaMP architecture in personalized news categorization.

In summary, this research establishes a robust methodology for personalized news categorization, providing valuable insights into the intricate interplay between diverse datasets, retrieval algorithms, and language models. This contribution significantly advances the field of tailored content recommendation systems, paving the way for more effective and personalized user experiences.

Keywords

Personalization; Large Language Models; Content Recommendation; LaMP Architecture; AG News Dataset; Book Depository Dataset; BM25 Algorithm;

1. INTRODUCTION

In the evolving landscape of Natural Language Processing (NLP), the pursuit of personalization has become a cornerstone, catering to user expectations for tailored and individualized experiences. Despite extensive exploration in various domains such as information retrieval and human-computer interaction, a comprehensive investigation of personalization within NLP remains relatively limited.

This research aligns with the architectural framework proposed by the LaMP paper [9], structured around the core concept of personalization for a given user-associated sample. The LaMP architecture incorporates three primary components: a query generation function (φq), a retrieval model (R), and a prompt construction function (φp). This framework aims to transform input into a cohesive prompt for a language model, facilitating personalized content recommendation.

In the pursuit of advancing personalization capabilities, this research explores how the LaMP architecture aligns with a news dataset, exemplified by the AG News dataset, and how it performs when applied to a distinct categorization domain—books—using the Book Depository dataset. The overarching objective is to discern whether a unified model can effectively categorize disparate data types, specifically news articles and books, which inherently share a similar nature.

The foundational premise involves a parallel architectural structure wherein the LaMP architecture is adeptly adapted and fine-tuned for the intricacies of a new personalization dataset. This dataset integration is twofold: first, with the AG News dataset, a comprehensive repository of news articles, and second, with the Book Depository dataset, a diverse collection of books. To facilitate document retrieval in both scenarios, the BM25 algorithm is employed, strategically selecting the top results and amalgamating them into prompts. These personalized prompts serve as the bedrock for a nuanced understanding of user preferences and enable effective content recommendation.

Simultaneously, the research acknowledges the growing significance of personalization in NLP and the challenges associated with limited user data. To address these challenges and lay the groundwork for a more comprehensive integration, the research takes a pioneering step by training the model on a books categorization dataset. This preliminary training enriches the model's understanding of semantic relationships, a strategic move aimed at enhancing the system's adaptability.

The integration of the LaMP architecture with both news and books datasets serves as a pivotal exploration into the potential universality of personalized categorization models. By assessing the model's performance across diverse datasets, the research seeks to uncover insights into its adaptability, robustness, and the extent to which it can serve as a unified solution for categorizing distinct yet thematically similar data types. The investigation, therefore, bridges a critical gap in existing literature by presenting a holistic approach to personalized text categorization, poised to transcend domain-specific limitations and cater to a spectrum of user interests.

As we delve deeper into the exploration of personalized text categorization, it is crucial to understand the intricate dynamics of the LaMP architecture when applied to different datasets. The AG News dataset, characterized by its rich repository of news articles spanning various categories, serves as an ideal testing ground for the LaMP architecture's adaptability in the news categorization domain. The performance metrics derived from this integration provide valuable insights into the architecture's effectiveness and potential enhancements in the context of news-related content.

Simultaneously, the research extends its inquiry into the realm of books categorization, leveraging the expansive Book Depository dataset. This dataset, encompassing a diverse collection of books across genres, poses a unique challenge and opportunity for the LaMP architecture. By exploring its performance in a distinct categorization domain, the study aims to uncover the architecture's versatility and its ability to transcend domain-specific constraints.

The document retrieval process, powered by the BM25 algorithm, plays a pivotal role in shaping the personalized prompts that drive the LaMP architecture. The strategic selection and amalgamation of top results from AG News and Book Depository datasets form the foundation for generating prompts that encapsulate user preferences. This iterative process, seamlessly integrating diverse datasets, contributes to a nuanced understanding of user-specific content consumption patterns.

The research also recognizes the inherent challenges associated with limited user data, prompting an innovative approach to preliminary training on a books categorization dataset. This strategic move aims to enrich the model's understanding of semantic relationships, equipping it with the adaptability required for diverse datasets. The amalgamation of this knowledge with the LaMP dataset accentuates the system's ability to overcome constraints related to data repetition and user-specific nuances observed in the LaMP news categorization dataset.

The investigation goes beyond the boundaries of traditional research by presenting a holistic approach to personalized text categorization. By embracing diversity in datasets and domains, the study lays the groundwork for a unified model capable of categorizing disparate yet thematically similar data types. The LaMP architecture, finely tuned and adapted for different datasets, emerges as a versatile solution, transcending domain-specific limitations and offering a comprehensive understanding of user preferences.

As the research unfolds, it opens avenues for future exploration and refinement. The comparative analysis of the LaMP architecture's performance in news and books categorization paves the way for a deeper understanding of its adaptability. Insights gained from this exploration can inform the development of more robust and universally applicable personalized content recommendation systems.

In conclusion, this research represents a significant stride towards unraveling the complexities of personalized text categorization. By integrating the LaMP architecture with diverse datasets, the study not only addresses current limitations but also propels the evolution of content recommendation systems towards greater adaptability. The quest for personalization in NLP takes a leap forward, offering a nuanced, contextually rich, and engaging content experience tailored to diverse user interests and preferences

2. Related Work

Several prior works have explored the realm of personalized content recommendation; however, a common limitation among them is the insufficient consideration of dataset diversity [1]. The study by Xue et al. [2], titled "Personalized Web Search by Mapping User Queries to Categories," primarily focuses on personalization strategies in web search. While the research contributes valuable insights into mapping user queries to specific categories for enhanced search experiences, it lacks a comprehensive exploration of diverse datasets. The absence of diverse datasets in this context might limit the adaptability and generalizability of the proposed strategies across various content domains.

Similarly, Naumov et al. [3] in "Enhancing Text Classification with Personalized Attention Mechanisms" delve into personalized attention mechanisms for text classification. Although the work explores tailored attention mechanisms to improve text classification, it falls short in addressing the impact of dataset diversity on the performance of these mechanisms. The reliance on specific datasets without considering a diverse array of content domains could limit the model's adaptability to broader contexts.

Research in the domain of dataset integration has witnessed significant contributions, yet the focus on diversity is often underemphasized. Rahm and Bernstein [4] in "A Survey of Integration Methods of Heterogeneous Databases" and Halevy et al. [5] in "Data Integration: The Teenage Years" provide valuable insights into heterogeneous database integration. However, these works predominantly address integration methodologies rather than emphasizing the inclusion of diverse datasets, a crucial factor for achieving a comprehensive understanding of user preferences.

In the document retrieval domain, studies such as Robertson et al. [6] in "Okapi at TREC-3" and Robertson and Zaragoza [7] in "Parameter-Free Probabilistic Document Ranking" have significantly influenced our approach. While these works lay the groundwork for effective document retrieval techniques, the lack of diversity in the evaluated datasets may limit the generalizability of the proposed algorithms across varied content domains.

Existing studies on language model finetuning, including Devlin et al. [8] in "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" and Flek [9] in "Leveraging Large Amounts of Weakly Supervised Data for Multi-Language Sentiment Classification," have been pivotal in advancing natural language processing. However, the potential impact of diverse datasets on improving model understanding and performance remains relatively unexplored in these works.

Our research aims to address these limitations by explicitly integrating diverse datasets, such as the LaMP News Categorization Dataset, AG News, and Book Depository Dataset [10]. The lack of diversity in prior research underscores the novelty and significance of our approach in leveraging varied content domains for personalized content recommendation. The inclusion of diverse datasets not only enhances the model's adaptability but also contributes to more inclusive and accurate personalized recommendations across different user preferences and interests.

3. Methodology

The methodology employed in this research focuses on achieving personalized content recommendation through the integration of

the LaMP architecture and a finetuned language model, specifically the Flan-T5-base-finetuned. The approach involves a meticulous data preprocessing pipeline and the utilization of both a new personalization dataset and existing datasets such as AG news and Book Depository. The following sections outline the detailed steps of the methodology.

In the pursuit of a highly personalized content recommendation system, we use the LaMP News Categorization Dataset. This dataset stands at the core of our methodology, serving as a foundation for understanding user preferences and generating personalized prompts. The LaMP News Categorization Dataset is designed to encapsulate a diverse array of news articles, each categorized into distinct topics.

For dataset diversification, we systematically integrate the LaMP News Categorization Dataset with the AG news dataset. This collaborative approach seeks to encompass a wide array of user interests, contributing to the refinement of our content recommendation system. The integration involves merging articles extracted from individual user profiles with the extensive collection of news articles present in the AG news dataset.

To address computational complexities and optimize processing efficiency, we judiciously curate a representative subset. This entails selecting a limited set, specifically opting for 100 documents from each distinct category within the AG news dataset. This pragmatic approach ensures not only a manageable dataset size but also preserves the richness and diversity of content, essential for generating effective and nuanced prompts.

The amalgamation of the LaMP News Categorization Dataset with the constrained yet diverse subset of the AG news dataset forms the foundation of our comprehensive dataset. This strategic integration lays the groundwork for a robust and nuanced understanding of user preferences, providing a detailed representation of their interests. Through this integration, we aim to strike a balance between computational efficiency and dataset richness, fostering a dataset that aligns with the intricacies of user-specific content consumption patterns.

Considering the extensive size of the Book Depository dataset, we thoughtfully select and curate three specific categories from this dataset. Additionally, to maintain alignment with the categories present in the LaMP dataset, we opt to rename the chosen categories. This procedural step is implemented to guarantee consistency and coherence. Importantly, the remaining preprocessing steps for this dataset closely mirror those applied to the AG news dataset.

The document retrieval process employs the BM25 algorithm, strategically designed to identify the top 5 documents that align most closely with each user's preferences. This methodology revolves around assigning scores to documents based on their similarity to user preferences, ultimately selecting the top candidates for recommendation. Notably, each instance of retrieving the top 5 documents involves personalizing the dataset pool according to the user's profile. In the case of AG News, we augment the dataset by appending articles from the user profile in the LaMP dataset, and similarly, for the Book Depository dataset. This personalized dataset construction is executed for every unique identifier within the LaMP dataset iteratively, every time we retrieve the top 5 documents.

The retrieved documents serve as the building blocks for prompt generation. A prompt is created in the string format: "The category for the article: {["text"]}+ is "+["category"]}". This formulation is applied to each of the top 5 documents. Subsequently, these prompts are concatenated into a single string.

Building upon the individual prompts for both the AG news and Book Depository datasets, a final prompt is constructed. This prompt takes the form: "[top 5 BM25 generated text]+ ‘The above are similar articles for the query asked further ‘+input." This comprehensive prompt encapsulates the user's preferences and sets the stage for the large language model.

I choose the Flan-T5-base-finetuned language model as the base. Its pre-trained capabilities and adaptability make it an ideal candidate for finetuning. The language model is finetuned using the final prompt generated. The model endeavors to predict the category of the input included in the prompt. This iterative process refines the model's understanding of user preferences and strengthens its ability to provide accurate content categorization.

To ensure seamless integration with the LaMP dataset, we carefully align the categories in the Book Depository dataset with those in the LaMP dataset. This harmonization step guarantees consistency and coherence in the model.

To assess the effectiveness of our model, we employ standard classification metrics such as accuracy and F1-score. These metrics provide a comprehensive understanding of the model's performance.

We undertake a comprehensive comparative analysis to assess the performance of our model under different integration scenarios. Specifically, we evaluate the model's performance when integrating the LaMP dataset with the news dataset and contrast it with its performance when integrated with the books dataset. This meticulous comparison is instrumental in elucidating the nuanced advancements facilitated by the LaMP architecture. By systematically examining these integration scenarios, we gain valuable insights into the model's efficacy and discern the impact of dataset integration on its overall performance.

4. Results

The training data allocated for LaMP news categorization underwent a meticulous split into training and validation subsets. Subsequently, the validation dataset was employed for thorough testing, leading to insightful results showcased in Table 1.

Table 1. This tables shows the results after evaluation for both the datasets

Dataset	Accuracy	F1 score
LaMP Dataset for news categorization + AG News Dataset	0.63403	0.572302
LaMP Dataset for news categorization + Books Depository Dataset	0.711026	0.667213

In the initial integration of the LaMP dataset for news categorization with the AG News dataset, the model achieved an accuracy of 0.63403 and an F1 score of 0.572302. Subsequently, when integrating with the Book Depository dataset, a notable improvement was observed, with accuracy increasing to 0.711026 and the F1 score rising to 0.667213.

The observed increase in both accuracy and F1 score during integration with the more diverse Book Depository dataset suggests a positive relationship between dataset diversity and model performance. This improvement underscores the model's adaptability to varied content domains, implying that a more diverse training dataset contributes to the development of robust models capable of accurately categorizing different types of content. The higher F1 score specifically highlights the model's ability to maintain a balance between precision and recall, indicating its effectiveness in correctly classifying instances while minimizing both false positives and false negatives.

These results affirm the potential advantages of training language models on datasets that represent diverse content domains. The model's ability to generalize across varied datasets enhances its accuracy and positions it as a versatile tool for categorizing content with diverse themes and characteristics, a crucial attribute for real-world applications where content spans multiple genres and domains.

5. Conclusion

In conclusion, our research has delved deeply into the intricate domain of personalized content recommendation, leveraging the LaMP architecture alongside a meticulously fine-tuned language model. The comprehensive integration of the LaMP dataset, centered on news categorization, with external datasets like AG News and Books Depository, has yielded a rich tapestry of insights.

Our study has brought forth compelling evidence, indicating a noteworthy improvement in both accuracy and F1 score through the judicious augmentation of training data with the Books Depository dataset. This substantial enhancement implies that broadening the model's exposure to a diverse array of content domains has a profoundly positive impact on its categorization prowess. The demonstrated versatility of the model positions it as a potentially versatile solution applicable across various data types, showcasing shared underlying patterns, as evident in the categorization of books and news articles.

Furthermore, the observed upswing in the F1 score emphasizes the model's adeptness in achieving a nuanced equilibrium between precision and recall. This crucial aspect not only fortifies its practical utility but also ensures accurate categorization while minimizing false positives and false negatives.

In essence, our research underscores the pivotal role played by dataset diversity in augmenting the overall effectiveness of personalized content recommendation systems. The promising outcomes achieved not only contribute to the current state of knowledge but also lay a robust foundation for future investigations and refinements in the realm of personalized Natural Language Processing (NLP) models. This marks a significant stride toward more nuanced, accurate, and tailored content recommendations, promising an enriched user experience across diverse domains. Our research serves as a catalyst for the evolution of personalized NLP, fostering the development of more resilient and adaptable content recommendation systems in the ever-evolving landscape of digital interactions.

6. Future Scope

The future trajectory of this research involves an expansive exploration of alternative retrieval models to comprehensively understand their impact on personalized content recommendation. Comparative analyses with state-of-the-art retrieval models will provide nuanced insights into the strengths and limitations of different approaches, fostering a deeper understanding of their applicability.

Additionally, there is a promising avenue for research in optimizing computational costs during the retrieval process. Exploring innovative techniques, such as advanced indexing methods or parallel processing, can potentially mitigate the computational burden associated with personalized document retrieval. This pursuit aligns with the overarching goal of refining and streamlining the recommendation system, making it more scalable and resource-efficient. Such endeavors are pivotal for the continual evolution of personalized NLP models, ensuring their feasibility and effectiveness in real-world, resource-constrained scenarios.

The future trajectory of this research involves an expansive exploration of alternative retrieval models to comprehensively understand their impact on personalized content recommendation. Comparative analyses with state-of-the-art retrieval models, including but not limited to those discussed in related works [1-9], will provide nuanced insights into the strengths and limitations of different approaches, fostering a deeper understanding of their applicability in diverse scenarios.

7. REFERENCES

[1] Xue, G., Zeng, H. J., Chen, Z., Yu, Y., & Ma, W. Y. (2009). Personalized web search by mapping user queries to categories. In Proceedings of the 18th international conference on World wide web (pp. 915-924).

[2] Naumov, M., Tavakol, M., & Dolog, P. (2019). Enhancing text classification with personalized attention mechanisms. In Proceedings of the 13th ACM Conference on Recommender Systems (pp. 198-206).

[3] Rahm, E., & Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. The VLDB Journal—The International Journal on Very Large Data Bases, 10(4), 334-350.

[4] Halevy, A., Ives, Z., Mork, P., Tatarinov, I., & Bernhardt, J. (2006). Data integration: The teenage years. In Proceedings of the 32nd international conference on Very large data bases (pp. 9-16).

[5] Robertson, S. E., Walker, S., Beaulieu, M., Gatford, M., & Payne, A. (1995). Okapi at TREC-3. In Proceedings of the third Text REtrieval Conference (TREC-3) (Vol. 95, pp. 109-126).

[6] Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. Foundations and Trends® in Information Retrieval, 3(4), 333-389.

[7] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[8] Flek, J. (2020). Leveraging large amounts of weakly supervised data for multi-language sentiment classification. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 6556-6566).

[9] Salemi, A., Mysore, S., Bendersky, M., & Zamani, H. (2021). LaMP: When Large Language Models Meet Personalization. arXiv preprint arXiv:2109.01381.

[10] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

[11] Baek, J., Chandrasekaran, N., & Cucerzan, S. (2020). Knowledge-augmented large language models for personalized contextual query suggestion. arXiv preprint arXiv:2010.02268.

[12] Lucie Flek. 2020. Returning the N to NLP: Towards contextually personalized classification models. In *Proceedings of the 58th Annual Meeting of the Asso- ciation for Computational Linguistics*, pages 7828– 7838, Online. Association for Computational Lin- guistics.