A COMPARATIVE STUDY OF LIBRARY CLASSIFICATION NUMBERS ASSIGNMENT ACCURACY: OPENAI LANGUAGE MODELS VS. HUMAN CLASSIFIERS

S. K. Illangarathne[1]

Abstract

This study examines the accuracy and feasibility of using OpenAI language models for assigning Dewey Decimal Classification (DDC) numbers in libraries. The primary objective was to compare their performance with that of human classifiers in accurately assigning DDC numbers. A sample of 30 books from various disciplines was used and the DDC numbers assigned by human classifiers serve as a benchmark tallying with the Online Catalogue of Library of Congress (OCLC Classify). The books were then presented to the linguaGPT language model for automated classification numbers. Statistical measures, such as precision, recall, and F1 score were used to evaluate the model's accuracy. The findings indicated that Open-AI language models demonstrate a moderate level of accuracy, with precision, recall, and F1 score of 0.53 (53%), suggesting they achieve moderate performance in correctly identifying positive instances and overall prediction accuracy. However, the model's performance varies depending on the complexity and specificity of the materials being classified. In comparison, human classifiers consistently achieve accurate classification, drawing on their expertise and contextual understanding. Recommendations from this study include: validating classification with reliable sources, seeking subject-matter expertise, regularly reviewing and updating the classification system and exploring hybrid AI-human systems. Implementing these recommendations can enhance the accuracy and reliability of library classification systems, facilitating improved access to information for users. In conclusion, OpenAI language models show promise in library classification, but improvements are needed to ensure greater accuracy.

Keywords: *Library classification, Number assignment accuracy, OpenAI language models, Human classifier, DDC*

[1] Senior Assistant Librarian, Faculty of Technology, Rajarata University of Sri Lanka.
E-mail: skillangarathne@gmail.com, isaman@rjt.ac.lk https://orcid.org/0000-0001-5331-2073

Introduction

The process of assigning accurate and consistent class numbers to books is a crucial aspect of library cataloguing systems. Traditionally, this task has been performed by human classifiers who possess expertise in Library and Information Science (LIS) and Classification Systems (CS) like the Dewey Decimal Classification (DDC) which is used in more than 135 countries and has been translated into over 30 languages (Comaromi, 1976; Liu, 1996). However, recent advancements in Natural Language Processing (NLP), specifically Open Artificial Intelligence (OpenAI) language models, have opened up possibilities for automated classification. Many researchers have indeed conducted numerous studies on OpenAI and its potential applicability to human activities. These studies have explored a wide range of topics, such as natural language processing, machine learning, computer vision, and reinforcement learning. Researchers have investigated the feasibility of using OpenAI models for tasks like language generation, translation, summarization, image recognition, automating and even game-playing (Gamage & Wanigasooriya, 2023).

Despite the potential benefits of using OpenAI language models for assigning DDC class numbers; there is a gap in research evaluating their accuracy and feasibility compared to human classifiers. This study aimed to fill that gap by conducting a comparative analysis of the class numbers generated by linguaGTP an OpenAI language model with those assigned by human classifiers in a real-world library context.

The significance of this research lies in its potential to revolutionize the process of assigning DDC class numbers in libraries. If proven to be accurate and reliable, OpenAI language models could present an efficient and cost-effective alternative to human classifiers. This could significantly streamline the cataloguing process, reduce human labour and improve the accessibility and discoverability of library resources.

By evaluating the performance of OpenAI language models against human classifiers, this study will provide empirical evidence regarding the capabilities, limitations and potential use of automated classification systems. The findings of this research will not only inform librarians and catalogers but also contribute to the broader discourse on the application of artificial intelligence in the field of library and information science.

Ultimately, this study aims to bridge the gap in knowledge by identifying whether OpenAI language models can be effectively integrated into library cataloguing processes, thereby creating opportunities for increased efficiency and accuracy in DDC class number assignment.

Objectives

The primary objective of this study was to assess the accuracy and feasibility of utilizing OpenAI language models for assigning Dewey Decimal Classification numbers in libraries.

Methodology

To obtain a representative sample, several University Library Online Public Access Catalogs (OPACs) were searched and crosschecked with OCLC Classify (https://classify.oclc.org/classify2/) for consistency and a total of 30 books written in the English language were randomly chosen from various disciplines. By analyzing the variations of the class numbers given by the individual classifiers and based on the authors' subject knowledge, assigned the idlest DDC class numbers for every book as human side output. After that, every above title was presented to a linguaGPT (Generative Pre-trained Transformer) version 4.0 for automated classification. Then model-generated DDC class numbers were compared with those assigned by human classifiers and tested their similarity and accuracy using statistical methods, such as precision, recall, and F1 score which were prominently used to evaluate the performance and reliability of the OpenAI language models in comparison to the human knowledge-based evaluations (Yacouby & Axman, 2020). Despite the various limitations associated with the evaluation methods mentioned above and the presence of unforeseen complexities throughout the entire process, the results were ultimately presented in a tabular format, accommodating these challenges.

The following formulas were used for precision, recall, and F1 score:

Precision: Precision measures the proportion of correctly identified instances in the predicted positive class.

Formula: Precision = True Positives / (True Positives + False Positives)

Recall: Recall measures the proportion of correctly identified instances in the actual positive class.

Formula: Recall = True Positives / (True Positives + False Negatives)

F1 score: The F1 score is a harmonic mean of precision and recall, providing a balanced measure that takes both metrics into account.

Formula: F1 score = 2 * (Precision * Recall) / (Precision + Recall)

In these formulas:

True Positives (TP) represents the number of correctly predicted positive instances.

False Positives (FP) represents the number of instances that are predicted positive but are negative.

False Negatives (FN) represent the number of instances that are predicted negative but are positive.

The above evaluation metrics are commonly used in binary classification tasks. They provide insights into the performance of a model or system by considering both the positive and negative predictions.

Results and Discussion

Table 1 illustrates the assigned DDC class numbers by the OpenAI language model and human classifiers for selected 30 book titles as an experiment. The results reveal OpenAI language models' ongoing challenge in assigning class numbers to somewhat complex titles that involve the use of "AND" ("Literature and Gender": Human DDC: 820.9 vs Modified DDC: 801.95) or titles encompassing more than two concepts ("Culture and Imperialism" = Human DDC: 809.93358 vs Modified DDC: 325.3).

Table 1: Assigned DDC Class Numbers by Human Classifiers and OpenAI Language Model

No	Title	Assigning DDC Human Classifiers	Class Numbers OpenAI Language Model
1	Concise Oxford Dictionary of Ecology	577.03	577.03
2	Modern Elementary Statistics	519.5	519.5
3	Essentials of Sociology: From Sociology--a Text with Adapted Readings	301	301
4	Nehru on World History	903	954.035
5	Literature and Gender	820.9	801.95
6	Guide to Surgical Appointments	610	617.9
7	Atkins' Physical Chemistry	541.3	541.2
8	Britain and Europe in the Seventeenth Century	942.2	941.06
9	Culture and Imperialism	809.93358	325.3

No	Title	Assigning DDC Human Classifiers	Class Numbers OpenAI Language Model
10	Oscar Wilde: The Critical Heritage	828.809	828.809
11	Principles of Macro-Economics	339	339
12	Cation Binding by Humic Substances	572.33	546.751
13	Foundations of Buddhism	294.3	294.3
14	Project Management: A Systems Approach to Planning, Scheduling, and Controlling	658.404	658.404
15	Macbeth	822.33	822.33
16	Essentials of Indian Philosophy	181.4	181.4
17	Photoshop CS4 All-in-One for Dummies	006.686	006.6
18	Barack Obama: The Story	973.932092	973.932
19	House of Doors	813.54	813.54
20	Seven Last Words	785.7194	232.9
21	Sustainable Agriculture	630	631.584
22	E-Commerce Systems Architecture and Applications	658.84	004.678
23	Introductory Econometrics: a Modern Approach	330.015195	330.015195
24	Russia-India-China: Evolution of Geo-political Strategic Trends	327.47054	320.12
25	Coffee Culture, Destinations and Tourism	306.4819	394.12
26	The Wealth of the Nation: An Economic History of the United States	330.973	330.9,
27	Computer-aided Design in Composite Material Technology III	620.118	620.193
28	The New Encyclopedia Britannica	031	030
29	The Handbook of Communication Skills	153.6	302.2
30	Molecular Biology and Biotechnology	574.88	572.8

Source: Survey Data, 2023

Table 2 displays precision, recall, and F1 score, all of which are recorded as 0.53 (53%). This suggests that the model or system has achieved a moderate level of performance in correctly identifying positive instances and overall prediction accuracy.

Table 2: Calculated Precision, Recall and F1 Score for the Human Classification vs. OpenAI Language Model Classification

Measure	How to measure?	Calculation	Rate
Precision	**Precision** = True Positives / (True Positives + False Positives) Eg: Out of 50 books, Model A accurately assigns 40 correct DDC numbers. This gives us a precision of 40/50 = 0.8 or 80%.	*16/30* = 0.5333	53%
Recall	Recall = True Positives / (True Positives + False Negatives) Eg: Out of 50 books correctly classified by Human A, Model A accurately assigns 40 correct DDC numbers. This gives us a recall of 40/50 = 0.8 or 80%.	*16 / (16 + 14)* *16/30* = 0.5333	53%
F1 score	2 * (Precision * Recall) / (Precision + Recall) Eg: The F1 score is the harmonic mean of precision and recall. If precision is 0.8 and recall is 0.8, then the F1 score is 2 * (0.8 * 0.8) / (0.8 + 0.8) = 0.8.	*2 x (0.53 x 0.53) / (0.53 + 0.53)* *2(0.2809/ 1.06)* *2 x 0.265* *= 0.53*	53%

Source: Survey Data, 2023

Conclusion and Recommendations

Based solely on these scores, it is challenging to draw specific conclusions or predictions about the model's capabilities or future performance. Further analysis and comparisons with other models or benchmarks are necessary to gain a more comprehensive understanding of the system's predictive accuracy and potential for improvement. However, it is noteworthy that, in most cases, the OpenAI language model has attempted to assign more accurate numbers but has not achieved 100% accuracy when compared to human catalogers. In a recent study conducted by Gamage and Wanigasooriya in 2023, similar findings were uncovered. They compared disparities in cataloguing frameworks by employing Anglo-American Cataloguing Rules 2 (AACR2) between AI systems and human experts. This underscores that, as a standalone solution, OpenAI may not achieve the same level of accuracy as a fully human approach for assigning class numbers to library books.

Another significant finding of the study is that the OpenAI language model struggles to provide precise classification numbers for books with complex scenarios when limited content details are available. However, based on the innovative aspects of the study, the following recommendations can be made when using OpenAI language models, not only for complex titles but also for simpler ones when assigning classification numbers:

1. Validate classification with reliable sources: Cross-verify classification using established library classification systems or professional librarians.
2. Seek subject-matter expertise: Engage experts to ensure accurate categorization of books in specific subject areas.
3. Regularly check updates in OpenAI language model and review the classification system

References

Comaromi, J. P. (1976). Knowledge organized is knowledge kept: The Dewey Decimal Classification, 1873-1976. *The Quarterly Journal of the Library of Congress, 33*(4), 311-331. https://www.jstor.org/stable/29781706

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542 (7639), 115-118.

Gamage, R., & Wanigasooriya, P. (2023). A study on the difference between AI-generated and human-expert-generated default cataloguing frameworks. In *International Research*

Conference of National Library of Sri Lanka – ICNATLIB 2023 (pp. 51-56), National Library and Documentation Services Board of Sri Lanka.

Liu, S. (1996). Decomposing DDC Synthesized Numbers. In 62nd IFLA General Conference - Conference Proceedings (pp. 299-308). Retrieved from http://archive.ifla.org/IV/ifla62/62-sonl.htm

McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. *AI Magazine*, 27(4), 12-14.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *In C.J. Burges and L. Bottou and M. Welling and Z. Ghahramani and K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems* (pp. 3111-3119). Curran Associates, Inc https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

Russell, S., & Norvig, P. (2010). Artificial Intelligence: A Modern Approach (3rd ed.). Prentice Hall.

Yacouby, R., & Axman, D. (2020). Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. *In Proceedings of the first workshop on evaluation and comparison of NLP systems* (pp. 79-91).