Beyond Traditional Training: GPT-Based Models in Zero-Shot Inference

Jia-Huei Liou

Abstract

This study explores whether GPT-based models surpass traditional logistic regression in zero-shot inference. A comparative analysis conducted on NLP tasks reveals that GPT models exhibit robust performance and adaptability under zero-shot conditions, matching or surpassing the efficacy of logistic regression models equipped with TF-IDF features. This paper highlights the potential of pretrained models to revolutionize machine learning strategies in NLP, emphasizing their ability to excel without task-specific training.

1 Introduction

In the rapidly advancing field of Natural Language Processing (NLP), the emergence of Generative Pretrained Transformer (GPT) models represents a significant leap forward. These models, when pretrained on extensive datasets, have demonstrated remarkable abilities across a wide range of NLP tasks. Notably, their capacity to engage in zero-shot learning — making accurate predictions on tasks without having been explicitly trained on them — offers a glimpse into a future where models adapt to new challenges with unprecedented flexibility. This stands in stark contrast to traditional machine learning approaches, such as logistic regression models that rely on Term Frequency-Inverse Document Frequency (TF-IDF) features, which necessitate task-specific data to function effectively.

Acknowledging that training GPT-based models on specific tasks could indeed enhance their performance further, this research intentionally focuses on zero-shot conditions. This decision stems from a desire to explore the foundational capabilities of GPT models in their "raw" form, without the advantages conferred by task-specific tuning. Such an investigation is vital for understanding the inherent strengths and limitations of these models, particularly in scenarios where collecting task-specific training data is impractical.

This paper makes several key contributions: It first undertakes a comparative analysis between the zero-shot capabilities of GPT-based models and the performance of traditional logistic regression models trained with TF-IDF features across NLP tasks. It then delves into the adaptability and efficiency of GPT models in interpreting and generating responses to previously unseen data. Finally, it contemplates the broader implications of these findings for the future of NLP applications, underscoring the transformative potential of pretrained models to alter machine learning strategies by diminishing the dependency on large, labeled datasets.

By emphasizing the zero-shot learning abilities of GPT-based models, this study aims to illuminate the shifting landscape of machine learning and NLP, offering insights into how these advanced models might be deployed in real-world applications where flexibility and rapid adaptability to new tasks are paramount.

2 Related work

The study "Unlocking Practical Applications in Legal Domain: Evaluation of GPT for Zero-Shot

Semantic Annotation of Legal Texts" by Jaromir Savelka investigates the use of GPT for semantic annotation in legal documents, achieving notable performance across various types of legal texts. This research underlines the utility of GPT models in legal contexts, such as document drafting and contract review, showcasing their potential to streamline complex legal workflows without task-specific training.

In parallel, "Multitask Prompted Training Enables Zero-Shot Task Generalization" by Victor Sanh et al. delves into zero-shot task generalization via explicit multitask learning. By fine-tuning a pretrained encoder-decoder model with a variety of prompted tasks, the study reveals that such models can excel in zero-shot performance, even outperforming significantly larger counterparts.

Both pieces of research emphasize the advanced adaptability and potential of pretrained models across both specialized domains and broader NLP challenges, contributing valuable insights to the field of machine learning and NLP.

3 Methods and materials

3.1 Datasets

In this study, a comprehensive dataset featuring various attributes of restaurants is utilized. This dataset includes customer-generated free-text reviews, numerical data such as review counts and star ratings, and categorical details like outdoor seating availability and state location. The dataset is divided into a training set, with restaurant types for our logistic regression model's training, and a test set, which omits these types. The primary focus is on analyzing customer reviews information using TF-IDF for logistic regression and zero-shot predictions with a GPT-based model. Comprising 13,144 reviews in the training set and 10,000 in the test set, this dataset provides a robust basis for our investigation. Additionally, the performance of models on the test data is evaluated through a Kaggle competition, offering a competitive and transparent platform to assess the effectiveness of our predictive approaches in real-world scenarios.

The distribution of restaurant types within the training dataset is presented in Table 1, detailing the frequency of each category.

Types	Count
american (traditional)	2,680
mexican	2,217
italian	2,032
chinese	1,696
american (new)	1,399
japanese	1,063
mediterranean	728
canadian (new)	484
thai	483
asian fusion	362

Table 1: The distribution of restaurant type in training dataset

3.2 Models

3.2.1 Logistic regression model

In this comprehensive study, we establish a logistic regression model as the foundational baseline against which we compare other methodologies, specifically focusing on the utilization of text preprocessing techniques and TF-IDF representation to prepare our dataset for analysis. The initial phase of our methodology involves an intricate text preprocessing routine designed to refine the raw textual data into a more analyzable format. This preprocessing includes several key steps: the removal of punctuation marks to reduce noise and prevent the model from misinterpreting

non-alphanumeric characters as part of the textual content; the tokenization of sentences into individual words to facilitate the analysis of the text at a granular level; the filtering out of stopwords—commonly used words in a language that offer little value to understanding the context of the text—as well as single characters, which are typically not informative and can skew the model's perception of important textual features; and the application of stemming, a process that reduces words to their root form, thereby allowing the model to recognize different forms of a word as a single entity. Subsequently, the TF-IDF vectorizer from scikit-learn is utilized to convert the processed text into a numerical format suitable for machine learning analysis. Through this detailed and systematic preparation of our data, leveraging both text preprocessing and TF-IDF vectorization, we lay a solid foundation for the logistic regression model to perform its analysis with a heightened ability to capture and interpret the essential characteristics of the textual data presented.

3.2.2 GPT-based model

For the zero-shot prediction component of this study, access to a GPT-based model is facilitated through the ChatGPT API, with the model of choice being GPT-3.5-turbo. This selection is grounded in the model's optimal balance between cost-efficiency and performance. Despite being a precursor to the more advanced GPT-4, GPT-3.5-turbo offers considerable advantages in terms of speed and affordability, making it well-suited for processing the extensive dataset of restaurant reviews without explicit training. This model's architecture, based on the transformer mechanism, is adept at handling complex language inputs, enabling nuanced understanding and generation of text. By crafting specific prompts that guide the model towards making informed predictions about restaurant types based solely on customer reviews, this research leverages the model's natural language processing capabilities to explore the potential of zero-shot learning in classifying diverse restaurant categories effectively.

3.3 Analyses

In the analyses of our logistic regression model, a nuanced technique was explored to enhance model accuracy by assigning increased weight to keywords indicative of certain restaurant types, such as "Chinese" for Chinese cuisine establishments. Although this method showed promise in improving performance, we chose to focus on a more straightforward approach for our baseline model to maintain simplicity and clarity in our comparative analysis.

For the GPT-based model predictions, a structured prompt was crafted and iteratively applied across the test dataset's reviews: “Review: {<the review in the test dataset>} \n\n Based on the above review, determine which restaurant style most match the review, among many restaurant styles only including [american (traditional), mexican, italian, chinese, american (new), japanese, mediterranean, canadian (new), thai, asian fusion], please just make one prediction.” Despite specific instructions, the GPT-based model occasionally predicted restaurant types not listed, such as Vietnamese or Korean. To address this, we queried the model for the closest match within the provided categories for any out-of-list predictions, adjusting the final prediction accordingly.

A significant challenge arose due to the rate limit of GPT-3.5-turbo, capped at 60,000 tokens per minute (TPM), leading to interruptions in the prediction process. This required restarting the model predictions every two minutes, adding to the time invested in the task. Moreover, the cost

associated with using the GPT-based model, approximately $4 for the analysis, presents a consideration for researchers and analysts looking to leverage such models for extensive review analysis. This cost, while modest, highlights the trade-offs between utilizing advanced NLP capabilities and managing resource constraints.

4 Results

Upon submitting the final predictions from the logistic regression baseline model and the GPT-based model to the Kaggle competition, I observed the following results, as detailed in Table 2:

Model	Private Score	Public Score
Logistic Regression Model	0.79768	0.79598
GPT-Based Model	0.80317	0.79497

Table 2: The result table showed in Kaggle competition with private score and public score

The comparison between the logistic regression model and the GPT-based model revealed intriguing outcomes. Notably, the GPT-based model slightly outperformed the logistic regression model in terms of the private score, indicating a marginally better generalization to unseen data. However, the logistic regression model showed a slightly superior public score, suggesting it may have better aligned with the public subset of the test data.

These results underscore the effectiveness of GPT-based models in zero-shot learning scenarios, demonstrating their capability to slightly surpass traditional logistic regression approaches in generalization, despite the logistic model's strong performance. The close scores between the two models highlight the competitive nature of NLP tasks and the potential of leveraging advanced pretrained models for complex prediction tasks without extensive task-specific training.

5 Discussion

The outcomes of our comparative analysis between the logistic regression baseline model and the GPT-based model provide valuable insights into the evolving landscape of natural language processing (NLP) and machine learning. The slightly higher private score achieved by the GPT-based model underscores the potential of advanced pretrained models to comprehend and categorize complex textual data effectively, even in a zero-shot learning context. This suggests that such models, with their vast knowledge bases and ability to infer from context, can offer a competitive edge in tasks requiring nuanced understanding of language.

Conversely, the logistic regression model's marginally better performance on the public score highlights the enduring relevance of traditional machine learning methods, particularly when optimized with techniques like TF-IDF vectorization and strategic weighting of indicative terms. This indicates that well-tuned traditional models remain a viable and efficient option for NLP tasks, especially in scenarios where computational resources or access to large-scale pretrained models might be limited.

The findings from this study contribute to the broader field of study in several ways. First, we can further explore the prompt variability. The performance of GPT-based models, especially in zero-shot learning scenarios, is significantly influenced by the design of the prompts used. Experimenting with various prompt formulations could reveal insights into how different models interpret and respond to instructions, potentially uncovering more effective ways to elicit accurate predictions. This experimentation might include variations in

prompt length, complexity, and specificity, offering a deeper understanding of the models' underlying language processing capabilities.

Furthermore, our study posits that enriching the dataset with a broader array of features could significantly refine the predictive accuracy of both logistic regression and GPT-based models. Currently centered on textual reviews, the expansion to include numerical ratings, categorical variables such as outdoor seating and state location, along with metadata like review timestamps, promises a comprehensive dataset analysis. Such an approach allows for a nuanced understanding of the data, potentially enhancing the models' classification capabilities regarding restaurant types.

For logistic regression models, advanced feature engineering and selection processes were applied to incorporate a diverse set of attributes, including customer reviews, restaurant names, ambience characteristics, geographical location, amenities like outdoor seating and Wi-Fi availability, and service options such as catering. The refined logistic regression model, bolstered by these additional features, exhibited notable improvements in performance, as evidenced by the results presented in Table 3. This underscores the value of a multifaceted data representation in boosting model efficacy.

Model	Private Score	Public Score
Logistic Regression Model with Features Engineering	0.82684	0.8415

Table 3: The result table showed in Kaggle competition with private score and public score

In the context of GPT-based models, the study explores the potential of augmenting prompt design to encapsulate these expanded dataset features, aiming to elevate the zero-shot learning capabilities. This adaptation seeks to harness the depth of context provided by the additional data points, thereby informing more accurate and contextually rich predictions.

This investigation into feature augmentation aligns with the broader objectives of enhancing NLP model performance and adaptability. By integrating a rich tapestry of dataset features, the study contributes to the ongoing discourse within the computational linguistics field, highlighting the importance of comprehensive data analysis and innovative model training approaches for the advancement of NLP methodologies.

6 Conclusion

This study embarked on a comparative analysis of logistic regression and GPT-based models to classify restaurant types based on customer reviews, with a particular focus on zero-shot learning capabilities. The investigation revealed that while both models exhibit strong performance, the GPT-based model demonstrated a slight edge in generalization ability under zero-shot conditions, as evidenced by its performance in the Kaggle competition. This finding underscores the potential of pretrained models to adapt and excel in tasks without explicit prior training, highlighting the significant advancements in NLP technologies.

The exploration into the augmentation of dataset features further emphasized the importance of comprehensive data representation in improving model performance. By integrating additional variables such as numerical ratings, categorical data, and metadata into the analysis, we were able to enhance the predictive accuracy of the logistic regression model significantly, thereby illustrating the value of feature engineering in traditional machine learning approaches.

In conclusion, this study contributes valuable insights into the capabilities and comparison of logistic regression and GPT-based models in classifying restaurant types from customer reviews. The findings advocate for a nuanced approach to model selection, emphasizing the importance of dataset features and prompting strategies in leveraging the full potential of NLP technologies. As the field continues to evolve, these insights will inform the development of more sophisticated, efficient, and adaptable NLP solutions.

References

[1] Jaromir Savelka. 2023. Unlocking Practical Applications in Legal Domain: Evaluation of GPT for Zero-Shot Semantic Annotation of Legal Texts. https://dl.acm.org/doi/abs/10.1145/3594536.3595161

[2] Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, Alexander M. Rush. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. https://arxiv.org/abs/2110.08207