Financial Sentiment Analysis on News and Reports Using Large Language Models and FinBERT

Yanxin Shen*
Zhejiang University
Hangzhou, Zhejiang, China
ssyysyx@zju.edu.cn

Pulin Kirin Zhang
Lehigh University
Bethlehem, Pennsylvania, USA
kirinpzhang@gmail.com

2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS) | 979-8-3503-7431-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICPICS62053.2024.10796670

***Abstract*—Financial sentiment analysis (FSA) is crucial for evaluating market sentiment and making well-informed financial decisions. The advent of large language models (LLMs) such as BERT and its financial variant, FinBERT, has notably enhanced sentiment analysis capabilities. This paper investigates the application of LLMs and FinBERT for FSA, comparing their performance on news articles, financial reports and company announcements. The study emphasizes the advantages of prompt engineering with zero-shot and few-shot strategy to improve sentiment classification accuracy. Experimental results indicate that GPT-4o, with few-shot examples of financial texts, can be as competent as a well fine-tuned FinBERT in this specialized field.**

Keywords—Sentiment Analysis, BERT, NLP, Large Language Models, Prompt Engineering, FinBERT

I. Introduction

Sentiment analysis has a long history [1], [2], [3] and focuses on identifying and extracting opinions and emotions from text data [4], [5]. It finds applications in numerous domains, including product reviews, social media posts, news articles, and financial reports, to glean valuable insights such as market trends, customer satisfaction, and the business performance of various entities [6], [7], [8], [9].

Grasping market sentiment is increasingly vital for those in the finance sector. Financial Sentiment Analysis (FSA) is an effective instrument for evaluating and interpreting market sentiment from textual data sources [7], [3], [10]. It delivers crucial insights into market trends, investor sentiment, and the potential effects of news and events on financial markets, enabling informed decision-making [11], [10], [11].

FSA presents several challenges that distinguish it from general sentiment analysis, such as the need for domain-specific knowledge, dealing with ambiguities, and managing uncertainties [12], [9]. To overcome these challenges, FSA can utilize the latest advancements in natural language processing (NLP) powered by large language models (LLMs) [13], [14], [15]. LLMs represent a new paradigm in Natural Language Understanding, setting them apart from traditional NLP models that depend on task-specific architectures, labeled data, and feature engineering [16]. LLMs offer numerous advantages over conventional NLP models, including scalability, generality, data efficiency, and transferability. However, they also have drawbacks, such as high computational costs, environmental impact, ethical concerns, and reliability issues.

Large language models (LLMs), including GPT-4 and Transformer-XL, epitomize the forefront of recent progress in natural language processing. These models excel in tasks like machine translation, sentiment analysis, and question-answering. They support various NLP applications and consistently push the boundaries of machine learning capabilities.

The few-shot strategy enables LLMs to achieve high accuracy even with limited examples, showcasing their potential for effective sentiment analysis in finance. Furthermore, this approach allows models to better understand nuanced financial language and context, ultimately leading to more reliable and insightful sentiment assessments. The novel concept of this research is to explore how prompting and LLMs can work together to handle the intricate and detailed data found in company financial reports within the financial domain.

Utilizing pre-trained knowledge and adapting it to specific domains and tasks is a major challenge for LLMs because of their training on extensive datasets. Prompt engineering, the craft of designing natural language input instructions known as prompts, offers a promising solution to this challenge. This technique does not require extra training or fine-tuning. By using regular language to interact with LLMs, prompt engineering enables them to learn effectively with or without prior examples.

Our research seeks to utilize prompt engineering for financial sentiment analysis classification within zero-shot and few-shot contexts, leveraging Large Language Models. The data is randomly extracted from financial news within the LexisNexis database. The objective is to meticulously design prompts that accurately capture sentiment categories and assist LLMs in understanding this sentiment task. Details on the concept of prompting are covered in Section II-D, and the specifics of the designed prompt are provided in Section III-B.

The primary aim is to evaluate how prompting handles complex data like financial reports and its capability to extract sentiment using large language models.

II. Related Work

Financial sentiment analysis covers a broad spectrum of methods and applications. This section offers a summary of the pertinent literature, categorizing it into three major areas.

A. Sentiment Analysis in Finance

Sentiment analysis entails extracting emotions or viewpoints from written text. Recent approaches can be categorized into two categories: 1) Methods in machine learning that derive features from text through the technique of "word counting", and 2) Deep learning methods that

represent text as a sequence of embeddings. The former struggles to capture the semantic information arising from specific word sequences, while the latter is often criticized for being "data-hungry" because of the vast number of parameters it must learn.

Financial sentiment analysis diverges from general sentiment analysis not only in its scope but also in its objectives, which is typically to predict market reactions to the data conveyed in the text. Loughran and McDonald (2016) present an extensive review of recent research on financial text analysis employing machine learning techniques such as "bag-of-words" or lexicon-based approaches.

Kraus and Feuerriegel (2017) were among the pioneers in applying deep learning techniques to analyze the polarity of financial texts. They employed an LSTM neural network to analyze ad-hoc company announcements, aiming to predict stock market movements. Their findings showed that this approach surpassed the performance of traditional machine learning methods. They determined that pre-training their model on a more extensive dataset yielded better results. However, they performed their pre-training on an annotated dataset. This approach is more restrictive compared to our approach of pre-training a language model using an unsupervised approach.

A wide range of studies make use of distinct neural architectures to conduct financial sentiment analysis. Sohangir et al. (2018) investigated several generic neural network architectures with a StockTwits dataset. They identified that CNN outperformed the other architectures. Lutz et al. (2018) employed doc2vec to create sentence embeddings from company-specific announcements. They also utilized multi-instance learning to forecast movements in the stock market. Maia et al. (2018) used text simplification alongside an LSTM network to categorize sentences from financial news by sentiment. This technique achieved leading results for the Financial PhraseBank, which is also employed in this thesis.

The scarcity of extensive labeled financial datasets poses a challenge in effectively utilizing neural networks for sentiment analysis. This limitation hinders the potential to fully exploit these models. Even when the initial word embedding layers are initialized with pre-trained values, the rest of the model still needs to learn complex relationships. This task becomes challenging due to the relatively limited amount of labeled data available. A potentially more effective method could be to initialize almost the entire model using pre-trained values. Afterward, these values can be fine-tuned specifically for the classification task. This approach may yield better results.

B. Text Classification with Pre-Trained Language Models

Language modeling is the process of predicting the next word in a text sequence. A notable recent breakthrough in natural language processing is the realization that models trained for language modeling can be efficiently fine-tuned for various downstream NLP tasks. This can be done with minimal modifications, enhancing their versatility and effectiveness across different applications. These models usually undergo training on large corpora. They are then fine-tuned on the target dataset by incorporating task-specific layers. The primary focus of this thesis is text classification, which serves as a clear example of this methodology. This approach is effectively demonstrated through text classification.

One of the early successful implementations of this technique was ELMo (Embeddings from Language Models). ELMo employs a deep bidirectional language model that is pre-trained on an extensive corpus. The hidden states of this model are used to generate a contextualized representation for each word. These pre-trained weights enable the computation of contextual word embeddings for any text. When utilized to initialize downstream tasks, these embeddings have demonstrated superior performance over static word embeddings like word2vec or GloVe in most tasks. For example, in text classification tasks such as SST-5, ELMo achieved state-of-the-art results when paired with a bi-attentive classification network.

C. Large Language Models

Large language models (LLMs) are trained using self-supervised learning goals. These models learn rich representations that allow them to understand syntactic, semantic, and pragmatic information. Moreover, LLMs can generate coherent and fluent natural language text, which can be applied to various downstream tasks.

LLMs differ from traditional NLP models and offer numerous advantages, such as scalability, generality, data efficiency, and transferability. The release of BERT in 2018 marked a major breakthrough for large language models. Since then, the number of language models available for various languages and domains has significantly increased.

OpenAI incorporated the transformer model in its releases of GPT1 in 2018, GPT2 in 2019, GPT3 in 2020, and ChatGPT in 2022. META introduced its LLMs, OPT and LLaMa. During the same period, BLOOM and Cohere1 were also developed.

By fine-tuning general-domain LLMs, financial domain models like BloombergGPT and FinGPT have emerged.

Newly released models include Google's BARD, LLaMa 2 [37], and GPT-4 integrated into Bing search. Nowadays, LLMs are being embedded in almost every website as chatbots, copilots, or AI assistants.

D. Prompt Engineering

Prompt engineering entails designing, refining, and optimizing prompts for generative AI systems capable of producing natural language outputs, such as text or graphics. This practice aids AI models in better comprehending questions and delivering more precise and relevant responses.

They have shown remarkable performance in numerous Natural Language Understanding (NLU) and Natural Language Generation (NLG) applications. Creating an effective and robust prompt necessitates careful design, considering format, length, style, and content. Evaluation methods should examine the quality, diversity, and consistency of responses, as well as the vulnerabilities of LLMs that could impact performance.

Prompt engineering can be performed using various techniques depending on the task's type and complexity, the available data's amount and quality, and the LLM's capabilities and limitations. Generally, these techniques are classified into zero-shot learning, few-shot learning,

instruction tuning, and prompt tuning.

With the advancements in LLMs and Conversational AI, prompt engineering is becoming a domain of its own. It is utilized for various tasks, such as sentiment analysis and humanlike summarization, achieved through carefully curated prompts. This research aims to analyze the effectiveness of prompting in financial sentiment analysis classification.

III. Methodology

This section outlines the methodology employed in this study, covering data collection through to outcome evaluation. The experiment aims to observe the response of LLMs to few-shot prompts in financial sentiment analysis.

A. Data Collection and Pre-processing

The primary dataset for sentiment analysis in this paper is the Financial PhraseBank from Malo et al. (2014). The Financial PhraseBank contains 4845 English sentences randomly selected from financial news in the LexisNexis database. These sentences were annotated by 16 individuals with expertise in finance and business. The annotators labeled the sentences based on how they believed the information might impact the stock price of the mentioned company.

The dataset also provides information on the level of agreement among annotators for each sentence. Table I shows the distribution of agreement levels and sentiment labels. To create a robust training set, we set aside 20% of the sentences as a test set and used 20% of the remaining sentences as a validation set. Our training set ultimately contains 3101 examples. For some experiments, we also employed 10-fold cross-validation.

The annotation process of the Financial PhraseBank ensures high-quality sentiment labels for the sentences, reflecting both the consensus and individual differences among multiple annotators. This diversity and consistency make the dataset highly reliable and practical for sentiment analysis research.For this research, the targeted data are financial reports of Pakistani companies. The director's review and chairman's review section hold the sentiment behind the report. These sections will be tested using prompt engineering on different LLMs.

Table I clearly represents the structure of the Financial PhraseBank dataset.

TABLE I. Financial PhraseBank Dataset

Item	***Description***
Data Source	Financial news and reports
Language	English
Number of Sentences	4,848
Classification Categories	Positive, Negative, Neutral, Uncertain
Annotation Method	Manually annotated by multiple financial professionals
Number of Examples per Category	- Positive: 1,484 - Negative: 846- Neutral: 1,953- Uncertain: 565
Application Areas	Financial sentiment analysis, Natural Language Processing (NLP)

B. Prompt Design

Following the principles of prompt design, we developed both zero-shot and few-shot prompts. Natural language is used to define the tasks for LLMs.

For zero-shot learning, we employed natural language queries, financial news, reports, and prompt questions to extract sentiment labels or classification results from the LLMs for input financial texts. The design and input prompt for the zero-shot setting are illustrated in Fig. 1.

system:

The goal is to classify financial texts (e.g., news articles, analyst reports, or company announcements) into three categories: positive, negative, or neutral sentiment. These sentiment categories are used to predict how the market might react to the financial information presented in the text.

user:

Please classify the following sentences into sentiment categories (0 for positive, 1 for negative, and 2 for neutral), and return a list of corresponding numbers, Only response the result list, do not say any word or explain.: {{ sentences_input }}

Fig. 1. Zero-shot prompt design

In a few-shot setting, multiple training examples are provided to help the LLMs better understand the task. Specifically, nine well-classified examples—comprising three positive, three negative, and three neutral news and reports with their ground truth labels—were included in the prompt design. Minor adjustments were made to the questions in this setting to give the LLMs more detailed guidance and rules. The design and input prompt for the few-shot setting are depicted in Fig. 2.

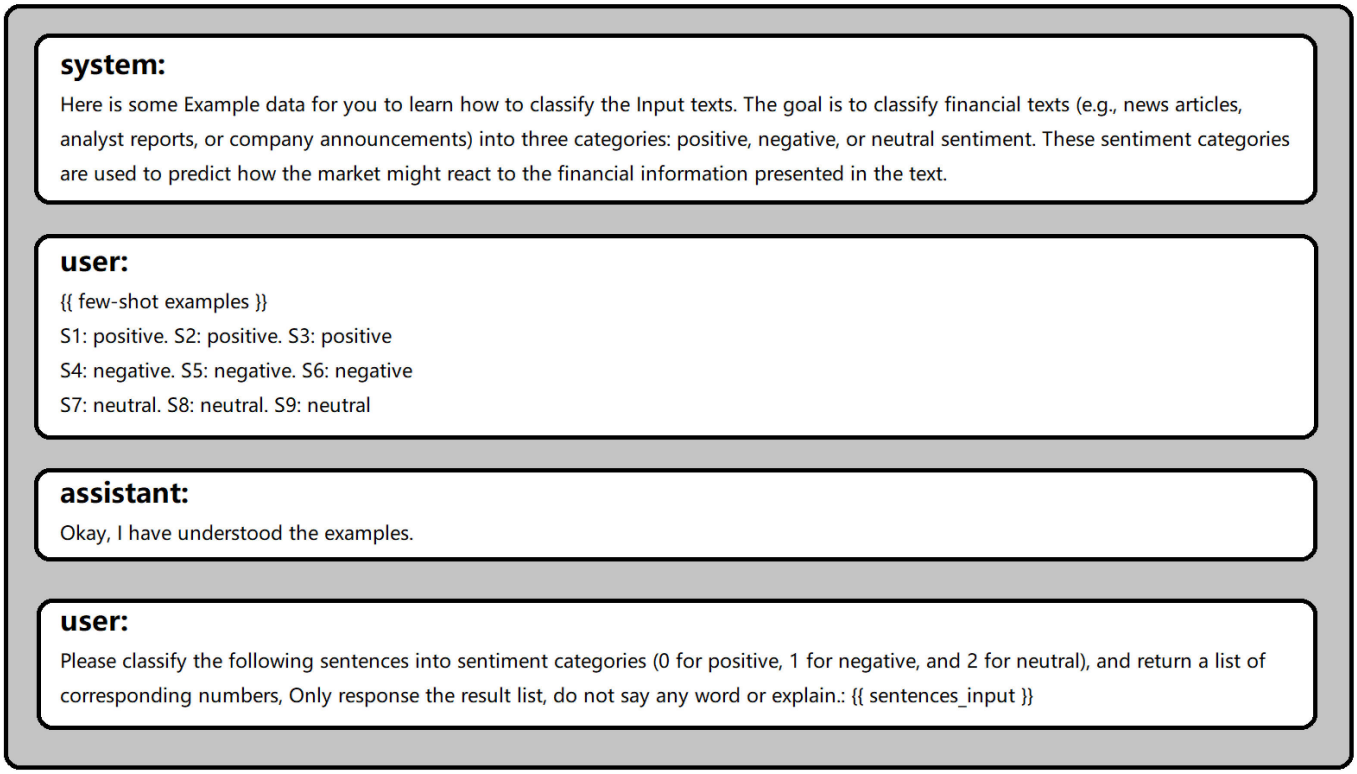

Fig. 2. Few-shot prompt design

C. LLMs Selection and Comparison Method

Typically, LLMs are chosen based on their size, availability, and task suitability. For this experiment, the two most commonly used and publicly accessible LLMs were selected: OpenAI ChatGPT (GPT 3.5) and OpenAI GPT 4. These conversational AI tools receive periodic updates, so the results may differ with future versions.

D. Fine-tuned BERT on Financial PhraseBank Dataset

We also use fine-tuned BERT on the financial sentiment analysis corpus of the size of 4848 labeled sentences and returned an accuracy of 0.88 in order to compare with LLMs. In FinBERT, BERT was adapted for the financial sector by additional pre-training on a financial dataset and fine-tuning for sentiment analysis. To the best of our knowledge, this study represents the first use of BERT in the finance field and is among the few that experimented with additional pre-training on a domain-specific corpus. The outcomes of FinBERT are presented in Table II.

IV. Experiments and Results

This section details the experimental setup and findings obtained from our financial sentiment analysis classification experiments. We utilized various models, including GPT-3.5-turbo, GPT-4o, and FinBERT, under different configurations included zero-shot, few-shot, and fine-tuning approaches. The primary objective was to evaluate the effectiveness of these models in accurately classifying financial sentiment from news articles and reports.

A. Experimental Setup

- Data Collection

The dataset used for this study is the Financial PhraseBank, which comprises 4,848 sentences extracted from financial news articles. Each sentence was labeled with one of four sentiment categories: Positive, Negative, Neutral, or Uncertain. The dataset was split into training, validation, and test sets to ensure a thorough evaluation. Specifically, 20% of the sentences were set aside as the test set, and another 20% of the remaining sentences were used for validation, leaving 3,101 sentences for training.

- Model Configurations

GPT-3.5-turbo and GPT-4o. Zero-shot: No prior examples are provided; the model must infer sentiment based on the query alone. Few-shot: A small number of examples are provided to help the model understand the task better.

FinBERT. Fine-tuned: The model underwent pre-training on a financial corpus, followed by fine-tuning on the Financial PhraseBank dataset for sentiment classification.

B. Results Analysis

Performance was assessed using four metrics: Accuracy, Precision, Recall, and F1-score. The results are presented in Table II.

TABLE II. Results on Financial PhraseBank Dataset

Model	Configuration	Metrics			
		Accuracy	*Precision*	*Recall*	*F1-score*
GPT-3.5-turbo	Zero-shot	0.78	0.79	0.84	0.80
	Few-shot	0.77	0.78	0.84	0.79
GPT-4o	Zero-shot	0.85	0.83	0.86	0.84
	Few-shot	0.86	**0.86**	0.84	0.85
FinBERT	Fine-tuned	**0.88**	0.85	**0.89**	**0.87**

C. Comparative Analysis

The experimental results indicate that fine-tuning FinBERT on a domain-specific corpus yields the best performance for financial sentiment analysis. The exceptional

performance of FinBERT can be linked to its pre-training on domain-specific data, which allows it to better understand the nuances of financial language. GPT-4o, with its advanced capabilities, also performed well, especially in the few-shot setting. The precision metric demonstrates that prompt engineering can greatly improve the efficiency of large language models even without extensive fine-tuning. Notably, the experimental results indicate that GPT-4o, with few-shot examples of financial texts, can be as competent as a well fine-tuned FinBERT in this specialized field. This finding underscores the potential of combining GPT-4o with prompt engineering techniques to achieve high performance in financial sentiment analysis. In conclusion, while GPT models show promise, FinBERT remains the most effective model for financial sentiment analysis due to its specialized training and fine-tuning on relevant financial data.

V. Conclusion

This study investigated the use of large language models (LLMs) and FinBERT for financial sentiment analysis (FSA) using financial news articles and reports. Our experiments assessed the performance of these models in various setups: zero-shot, few-shot, and fine-tuning.

GPT-3.5-turbo and GPT-4o demonstrated notable performance improvements with prompt engineering techniques, particularly in few-shot settings. However, their accuracy and overall effectiveness varied based on the complexity and specificity of the prompts. FinBERT, which was fine-tuned on the Financial PhraseBank dataset, consistently outperformed general-purpose LLMs. It achieved the highest scores in accuracy, precision, recall, and F1-score. These results highlight the significance of domain-specific pre-training for effective financial sentiment analysis.

Prompt engineering significantly enhanced the performance of LLMs. Few-shot prompts, in particular, provided better context, leading to more accurate and nuanced sentiment classifications. Zero-shot settings, while effective to some extent, often lacked the necessary context for precise sentiment extraction, highlighting the limitations of relying solely on general language models without fine-tuning.

The superior performance of FinBERT emphasizes the value of domain-specific models in specialized fields such as finance. FinBERT's ability to understand financial jargon and context-specific nuances makes it a robust tool for FSA since its development in 2019.

However, experimental results indicate that GPT-4o, with few-shot examples of financial texts, can achieve performance in financial sentiment analysis as competent as FinBERT. On a side note, the study revealed challenges in achieving high accuracy with zero-shot learning due to the lack of contextual information.

Future research should focus on improving prompt designs and incorporating more comprehensive few-shot examples to bridge this gap. Further exploration into fine-tuning strategies and leveraging additional domain-specific datasets could enhance the capabilities of LLMs for FSA.There is also potential for integrating LLMs with real-time financial data, enabling dynamic and context-aware sentiment analysis that can adapt to market fluctuations and emerging trends.

In conclusion, while domain-specific models like FinBERT remain superior due to their tailored pre-training, general-purpose LLMs like GPT-3.5-turbo and GPT-4o show promise for financial sentiment analysis with little setup. Effective prompt engineering can significantly enhance the performance of LLMs, making them viable tools for FSA in real-world applications. Future work should continue to refine these models and explore innovative approaches to further improve their accuracy and contextual understanding.

References

[1] Z. Nanli, Z. Ping, L. Weiguo, and C. Meng, "Sentiment analysis: A literature review," in 2012 International Symposium on Management of Technology (ISMOT), pp. 572–576, IEEE, 2012.

[2] M. V. Mäntylä, D. Graziotin, and M. Kuutila, "The evolution of sentiment analysis—a review of research topics, venues, and top cited papers," Computer Science Review, vol. 27, pp. 16–32, 2018.

[3] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: a review," Artificial Intelligence Review, vol. 53, no. 6, pp. 4335–4385, 2020.

[4] S. Liu, C. Zheng, O. Demasi, S. Sabour, Y. Li, Z. Yu, Y. Jiang, and M. Huang, "Towards emotional support dialog systems," arXiv preprint arXiv:2106.01144, 2021.

[5] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, "Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research," IEEE Transactions on Affective Computing, 2020.

[6] F. Xing, L. Malandri, Y. Zhang, and E. Cambria, "Financial sentiment analysis: An investigation into common mistakes and silver bullets," in Proceedings of the 28th International Conference on Computational Linguistics, (Barcelona, Spain (Online)), pp. 978‑987, International Committee on Computational Linguistics, Dec. 2020.

[7] X. Man, T. Luo, and J. Lin, "Financial sentiment analysis (fsa): A survey," in 2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS), pp. 617–622, IEEE, 2019.

[8] F. Barbieri, J. Camacho-Collados, L. Neves, and L. Espinosa-Anke, "Tweeteval: Unified benchmark and comparative evaluation for tweet classification," arXiv preprint arXiv:2010.12421, 2020.

[9] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," Artificial Intelligence Review, vol. 55, no. 7, pp. 5731–5780, 2022.

[10] F. Z. Xing, E. Cambria, and R. E. Welsch, "Natural language based financial forecasting: a survey," Artificial Intelligence Review, vol. 50, no. 1, pp. 49–73, 2018.

[11] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al., "Sparks of artificial general intelligence: Early experiments with gpt-4," arXiv preprint arXiv:2303.12712, 2023.

[12] H. Xu, B. Liu, L. Shu, and P. S. Yu, "Dombert: Domain-oriented language model for aspect-based sentiment analysis," arXiv preprint arXiv:2004.13816, 2020.

[13] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al., "A survey of large language models," arXiv preprint arXiv:2303.18223, 2023.

[14] L. Fan, L. Li, Z. Ma, S. Lee, H. Yu, and L. Hemphill, "A bibliometric review of large language models research from 2017 to 2023," arXiv preprint arXiv:2304.02020, 2023.

[15] D. Araci, "Finbert: Financial sentiment analysis with pre-trained language models," arXiv preprint arXiv:1908.10063, 2019.

[16] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," ACM Computing Surveys, vol. 55, no. 9, pp. 1–35, 2023.