CryptoLLM: Unleashing the Power of Prompted LLMs for SmartQnA and Classification of Crypto Posts

Aniket Deroy[1,*,†], Subhankar Maity[1]

[1]*IIT Kharagpur, Kharagpur, India*

Abstract

The rapid growth of social media has resulted in an large volume of user-generated content, particularly in niche domains such as cryptocurrency. This task focuses on developing robust classification models to accurately categorize cryptocurrency-related social media posts into predefined classes, including but not limited to objective, positive, negative, etc. Additionally, the task requires participants to identify the most relevant answers from a set of posts in response to specific questions. By leveraging advanced LLMs, this research aims to enhance the understanding and filtering of cryptocurrency discourse, thereby facilitating more informed decision-making in this volatile sector. We have used a prompt-based technique to solve the classification task for reddit posts and twitter posts. Also, we have used 64-shot technique along with prompts on GPT-4-Turbo model to determine whether a answer is relevant to a question or not.

Keywords

GPT, Relevance, Classification, Few-shot, Prompt Engineering

1. Introduction

We have seen the emergence of cryptocurrencies has generated substantial interest and discourse across various social media platforms [1]. With millions of users expressing their opinions, sharing information, and speculating on the future of digital assets, the need for effective tools to process and analyze this data has become increasingly critical [2]. However, the diverse nature of social media content, characterized by informal language, varying tones, and mixed sentiments, presents significant challenges for traditional text classification methods [3].

This task is designed to address these challenges by inviting participants to develop machine learning models capable of accurately classifying cryptocurrency-related social media posts into one of eight categories, such as objective, positive, negative, and others. Beyond simple classification, the task also extends to the identification of relevant answers to specific questions from a corpus of posts, a function that mirrors real-world applications where users seek precise information in a sea of data [4].

By focusing on the cryptocurrency domain, this task [5] not only contributes to the broader field of sentiment analysis and text classification but also offers practical implications for stakeholders in the financial sector. Improved classification and information retrieval from social media can lead to better market insights, more effective communication strategies, and enhanced user experiences in digital finance. This study aims to showcase the potential of natural language processing techniques in transforming raw social media data into actionable intelligence in the context of cryptocurrency [6]. In our approach to solving the classification task for Reddit and Twitter posts, we employed a prompt-based technique [7], which has proven to be effective in leveraging the strengths of large language models for text classification. Specifically, we utilized carefully designed prompts to guide the model in understanding the context and nuances of social media posts related to cryptocurrency. This approach enabled us to harness the power of GPT-4-Turbo [8] to classify posts into predefined categories, such as objective, positive, negative, etc.

To enhance the model's ability to determine the relevance of answers to specific questions, we

implemented a 64-shot technique [9] in conjunction with our prompt-based method to determine whether a answer is relevant to a question or not. This allowed the model to learn from a variety of examples, thereby improving its accuracy in identifying whether a given answer is pertinent to the corresponding question.

2. Related Work

The task of classifying and analyzing social media posts, particularly in specialized domains such as cryptocurrency, has garnered significant attention in recent years [3]. Social media platforms like Twitter and Reddit are rich sources of user-generated content that reflect public sentiment, market trends, and community discussions [10]. However, the unstructured and often noisy nature of social media data presents challenges for traditional text classification and information retrieval methods [11]. Early efforts in text classification on social media data primarily relied on traditional machine learning approaches, such as support vector machines (SVMs) and logistic regression, combined with handcrafted features like term frequency-inverse document frequency (TF-IDF) and n-grams [12, 13, 14, 15]. However, these approaches often struggled with the informal and context-dependent language prevalent on platforms like Twitter and Reddit.

With the advent of deep learning, particularly with the development of convolutional neural networks (CNNs) and recurrent neural networks (RNNs), the field saw a shift towards more robust models that could capture the nuances of language more effectively [16, 17, 18, 19, 20]. More recently, transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) have set new benchmarks in text classification by leveraging contextual embeddings to better understand the intricacies of language [21, 22]. These models have been particularly effective in sentiment analysis, where understanding the context and tone of a post is crucial.

Prompt-based learning [7], as popularized by models such as GPT-3 [9] and GPT-4-Turbo [8], has emerged as a powerful paradigm in NLP. Unlike traditional supervised learning approaches, prompt-based methods guide the model's predictions by framing the task as a completion or question-answering problem. This technique has been shown to be highly effective, particularly when dealing with limited labeled data, as it allows the model to leverage its pre-existing knowledge more effectively . In our work, we build on this approach by designing specific prompts tailored to the cryptocurrency domain, enhancing the model's ability to classify posts and identify relevant information.

Few-shot learning techniques [9] have gained traction as a means to improve model performance in scenarios where labeled data is scarce. In particular, the use of n-shot learning, where the model is exposed to a small number of labeled examples during training, has shown promise in various NLP tasks [23]. The 64-shot technique we employed is an extension of this concept, allowing the model to learn from a diverse set of examples, thereby improving its generalization to unseen data . Recent studies have demonstrated the effectiveness of few-shot learning in improving the performance of large language models (LLMs) on various tasks, including text classification [24, 25, 26] and question answering [27, 28].

Research specifically targeting cryptocurrency-related social media content has been limited but is growing in importance as digital currencies become more mainstream. Previous studies have primarily focused on sentiment analysis [16, 17], attempting to correlate social media sentiment with cryptocurrency price movements . However, there has been a lack of comprehensive studies that address both classification and relevance determination in this domain. Our work contributes to filling this gap by combining prompt-based learning and few-shot techniques to handle the unique challenges presented by cryptocurrency-related social media posts. In summary, our work builds on the advancements in LLMs, prompt engineering, and few-shot learning to address the specific challenges of classifying and retrieving relevant information from cryptocurrency-related social media data. By integrating these techniques, we aim to push the boundaries of what can be achieved in this specialized and dynamic field.

3. Dataset

The QnA task has 6323 posts in the test set. The Twitter opinion dataset has 500 posts in the test set. The Reddit opinion dataset has 500 posts in the test set.

We use the reddit opinion dataset in Zero-shot setting. We use the twitter opinion dataset in Zero-shot setting. We use the QnA dataset in 64-shot setting. Since we primarily use only the test set data for our predictions we only mention the statistics of the test set data.

4. Task Definition

The Task-1 is to develop a classification model to classify cryptocurrency related social media posts into eight classes namely, *Noise, Objective, Positive, Negative, Neutral, Question, Advertisement, Miscellaneous.*

The Task-2 is, for a given question wrt cryptocurrency related post we have to detect all the answers which are relevant to this question.

5. Methodology

5.1. Why Prompting?

- **Leveraging Pretrained Language Model Knowledge:** Prompting allows us to directly utilize the extensive knowledge embedded in large pre-trained language models like GPT-4-Turbo [7]. These models are trained on vast and diverse datasets, capturing a wide range of linguistic patterns, contextual nuances, and domain-specific information. By crafting appropriate prompts, we can guide the model to apply this knowledge effectively to specific tasks, such as classifying social media posts or finding relevant answers [7].

- **Reducing the Need for Large Labeled Datasets:** Traditional machine learning approaches often require substantial labeled datasets to achieve high performance [29]. However, obtaining large, high-quality labeled datasets, especially in specialized domains like cryptocurrency, can be resource intensive. Prompting, especially when combined with few-shot learning techniques, mitigates this need by enabling the model to perform well with minimal labeled examples [30]. This is particularly beneficial in scenarios where labeled data is scarce or costly to acquire.

- **Flexibility and Adaptability:** Prompting offers significant flexibility in task formulation. By simply changing the structure or wording of a prompt, we can adapt the model to perform a wide variety of tasks without the need for retraining [7]. This adaptability is crucial when dealing with dynamic content, such as social media posts, where topics, language, and context can rapidly evolve.

- **Improved Contextual Understanding:** Social media posts, especially those related to cryptocurrency, often contain informal language, abbreviations, slang, and domain-specific jargon [31]. Traditional classification models might struggle with this variability. Prompt-based approaches, however, leverage the model's deep contextual understanding, allowing it to better interpret and classify such content [7]. For instance, a well-crafted prompt can help the model distinguish between positive and negative sentiment even when the language is complex or non-standard.

- **Enhanced Relevance Determination:** Finding relevant answers to specific questions within a corpus of social media posts is a complex task, as it requires understanding the subtle relationships between questions and potential answers [32]. Prompting allows us to frame this task in a way that aligns with the model's strengths, treating it as a natural language inference problem where the model evaluates the likelihood that a given answer is relevant to the question [7]. Using prompt-based few-shot learning, we can further fine-tune this ability, improving the accuracy of relevance determination in contexts where direct answer-question pairs are not always straightforward [28].

- **Rapid Prototyping and Experimentation:** Prompt-based techniques facilitate rapid prototyping and experimentation, allowing researchers and practitioners to quickly test different approaches to a problem [33]. This is particularly useful in the fast-paced world of cryptocurrency, where market conditions and public sentiment can change rapidly [34]. Prompting enables quick adjustments to the behavior of the model, ensuring that it remains effective even as the underlying data evolves [7].
- **Alignment with Human-Like Reasoning:** Prompt-based methods are designed to mimic how humans might approach a task, framing it naturally and intuitively [35]. This alignment with human-like reasoning makes the model's predictions more interpretable and reliable, particularly in tasks like classification and relevance determination where understanding subtle nuances is crucial [36]. By asking the model to "complete" or "answer" in a way that mirrors human thought processes, we can achieve more accurate and contextually appropriate outcomes.

In summary, prompting offers a powerful and efficient way to solve classification tasks and find relevant answers in complex, dynamic, and data-scarce environments like cryptocurrency-related social media. It capitalizes on the strengths of large language models, reducing the need for extensive labeled data, and providing flexibility, adaptability, and improved performance across a wide range of tasks.

5.2. Prompt Engineering along with Few-shot Methods

We used the GPT-4-Turbo model via prompting through the OpenAI API[1] to solve the classification task. After the prompt is provided to the LLM, the following steps happen internally while generating the output. We outline the steps that occur internally within the LLM (i.e., a summary of the prompting approach using GPT-4 Turbo).

Step 1: Tokenization

- **Prompt:** $X = [x_1, x_2, \ldots, x_n]$
- The input text (prompt) is first tokenized into smaller units called tokens. These tokens are often subwords or characters, depending on the model's design.
- **Tokenized Input:** $T = [t_1, t_2, \ldots, t_m]$

Step 2: Embedding

- Each token is converted into a high-dimensional vector (embedding) using an embedding matrix E.
- **Embedding Matrix:** $E \in \mathbb{R}^{|V| \times d}$, where $|V|$ is the size of the vocabulary and d is the embedding dimension.
- **Embedded Tokens:** $T_{\text{emb}} = [E(t_1), E(t_2), \ldots, E(t_m)]$

Step 3: Positional Encoding

- Since the model processes sequences, it adds positional information to the embeddings to capture the order of tokens.
- **Positional Encoding:** $P(t_i)$
- **Input to the Model:** $Z = T_{\text{emb}} + P$

Step 4: Attention Mechanism (Transformer Architecture)

- **Attention Score Calculation:** The model computes attention scores to determine the importance of each token relative to others in the sequence.

---

[1] https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4

- **Attention Formula:**

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

- where Q (query), K (key), and V (value) are linear transformations of the input Z.
- This attention mechanism is applied multiple times through multi-head attention, allowing the model to focus on different parts of the sequence simultaneously.

Step 5: Feedforward Neural Networks

- The output of the attention mechanism is passed through feedforward neural networks, which apply non-linear transformations.
- **Feedforward Layer:**

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{2}$$

- where W_1, W_2 are weight matrices and b_1, b_2 are biases.

Step 6: Stacking Layers

- Multiple layers of attention and feedforward networks are stacked, each with its own set of parameters. This forms the "deep" in deep learning.
- **Layer Output:**

$$H^{(l)} = \text{LayerNorm}(Z^{(l)} + \text{Attention}(Q^{(l)}, K^{(l)}, V^{(l)})) \tag{3}$$

$$Z^{(l+1)} = \text{LayerNorm}(H^{(l)} + \text{FFN}(H^{(l)})) \tag{4}$$

Step 7: Output Generation

- The final output of the stacked layers is a sequence of vectors.
- These vectors are projected back into the token space using a softmax layer to predict the next token or word in the sequence.
- **Softmax Function:**

$$P(y_i|X) = \frac{\exp(Z_i)}{\sum_{j=1}^{|V|} \exp(Z_j)} \tag{5}$$

- where Z_i is the logit corresponding to token i in the vocabulary.
- The model generates the next token in the sequence based on the probability distribution, and the process repeats until the end of the output sequence is reached.

Step 8: Decoding

- The predicted tokens are then decoded back into text, forming the final output.
- **Output Text:** $Y = [y_1, y_2, \ldots, y_k]$

The prompt for classifying Twitter posts: *"Classify the Twitter post <Twitter_post> into one of the following labels: Noise, Objective, Negative, Positive, Neutral, Question, Advertisement, Miscellaneous"*. The overview diagram for the classification of Twitter posts is presented in Figure 1.
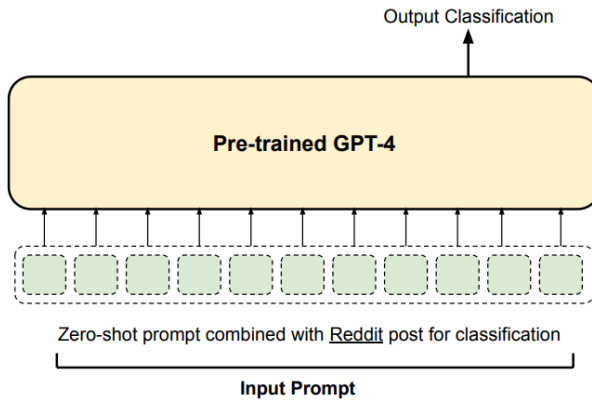
The prompt for classifying Reddit posts: *"Classify the Reddit post <Reddit_post> into one of the following labels: Noise, Objective, Negative, Positive, Neutral, Question, Advertisement, Miscellaneous"*. The overview diagram for the classification of Reddit posts is depicted in Figure 2.

Along with the 64-shot example, the following prompt will be provided, *"Given the title <title> and comment <comment_body>. Please check whether the comment is relevant or not-relevant to the title. Only state relevant or not-relevant."* The overview diagram for determining relevance of posts is shown in Figure 3.
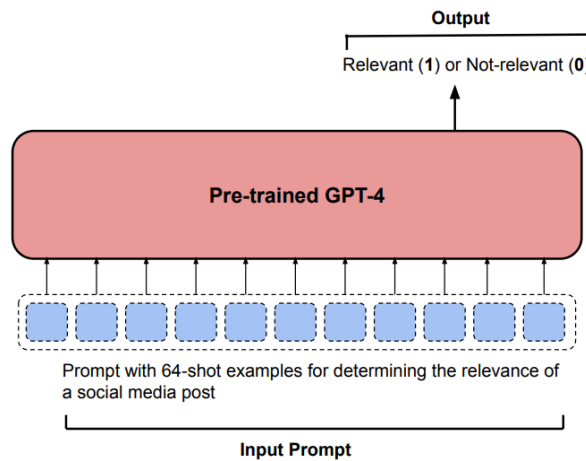
For the results reported, we ran the GPT-4-Turbo model at a temperature of 0.7.
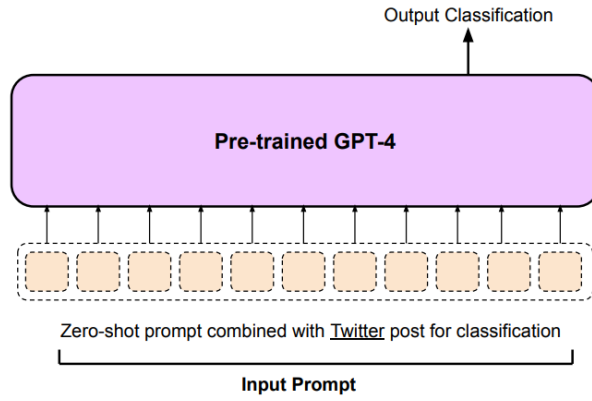
Figure 1: An overview of the methodology for classifying Twitter posts using GPT-4-Turbo.

Figure 2: An overview of the methodology for classifying Reddit posts using GPT-4-Turbo.

Figure 3: An overview of the methodology for determining the relevance of a social media post using GPT-4-Turbo.

6. Results

For our team named **TextTitans**, the Macro-F1 score for reddit opinion task is 0.249 with a rank of 8th. A score of 0.249 indicates that the model's performance is relatively low, possibly due to challenges in handling the specific data or class imbalances in this task.

The macro-F1 score for twitter opinion task is 0.266 with a rank of 8th. The slightly higher score compared to the Reddit task suggests that the model performed marginally better on Twitter data but still struggled overall.

The macro-F1 for QnA task is 0.157 with a rank of 1st. For the QnA task, the model achieved a Macro-F1 score of 0.157, which is the lowest score among the three tasks. However, the model ranked 1st in this task, indicating that, despite the low score, it outperformed other models in this specific challenge. This could suggest that the QnA task is particularly difficult, leading to lower scores overall, but the model was still the best among its competitors.

7. Conclusion

This study evaluated the performance of our model across three distinct tasks: Reddit opinion classification, Twitter opinion classification, and Question-and-Answer (QnA) task. The Macro-F1 scores for the Reddit and Twitter opinion tasks were 0.249 and 0.266, respectively, placing the model 8th in both tasks. These results highlight the challenges our model faced in effectively handling the varied and complex nature of social media data, possibly due to class imbalances or the nuanced expressions of opinions on these platforms.

In contrast, the model achieved a Macro-F1 score of 0.157 in the QnA task, securing the 1st position. Despite the lower absolute score, this top ranking indicates that our model outperformed others in this particularly challenging task, suggesting its robustness in scenarios where the data complexity is high and the overall performance across models is lower.

These findings underscore the importance of tailoring models to the specific characteristics of each task and dataset. Future work will focus on refining the model to better capture the intricacies of social media language and enhancing its ability to handle a broader range of tasks with more balanced performance across all metrics.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to: Drafting content, Grammar and spelling check, etc. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] H. N. D. Şenyapar, Cryptocurrency on social media: Analyzing the digital discourse towards the coin market, İktisadi İdari ve Siyasal Araştırmalar Dergisi 9 (2024) 202–223.

[2] W. Sutherland, M. H. Jarrahi, The sharing economy and digital platforms: A review and research agenda, International Journal of Information Management 43 (2018) 328–341.

[3] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, A. Hussain, Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions, Information Fusion 91 (2023) 424–444.

[4] L. Nizzoli, M. Avvenuti, M. Tesconi, S. Cresci, R. Di Pietro, W. Quattrociocchi, F. Gini, Leveraging Social Media and AI to foster Secure Societies against Online and Offline Threats., Ph.D. thesis, University of Pisa, Italy, 2021.

[5] S. Sarkar, G. Sen, K. Rudra, K. Ghosh, Understanding cryptocurrency related opinions and questions from social media posts (cryptoqa 2024), in: Forum of Information Retrieval (FIRE), 2024. URL: https://sites.google.com/view/cryptoqa-2024.

[6] K. I. Roumeliotis, N. D. Tselikas, D. K. Nasiopoulos, Llms and nlp models in cryptocurrency sentiment analysis: A comparative classification study, Big Data and Cognitive Computing 8 (2024) 63.

[7] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Computing Surveys 55 (2023) 1–35.

[8] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[9] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[10] Y. Zhan, R. Liu, Q. Li, S. J. Leischow, D. D. Zeng, Identifying topics for e-cigarette user-generated contents: a case study from multiple social media platforms, Journal of medical Internet research 19 (2017) e24.

[11] R. Irfan, C. K. King, D. Grages, S. Ewen, S. U. Khan, S. A. Madani, J. Kolodziej, L. Wang, D. Chen, A. Rayes, et al., A survey on text mining in social networks, The Knowledge Engineering Review 30 (2015) 157–170.

[12] K. Y. Tai, J. Dhaliwal, S. M. Shariff, Online social networks and writing styles–a review of the multidisciplinary literature, Ieee Access 8 (2020) 67024–67046.

[13] K. Taha, P. D. Yoo, C. Yeun, A. Taha, Text classification: A review, empirical, and experimental evaluation, arXiv preprint arXiv:2401.12982 (2024).

[14] M. Errami, M. A. Ouassil, R. Rachidi, B. Cherradi, S. Hamida, A. Raihani, Sentiment analysis onmoroccan dialect based on ml and social media content detection (2023).

[15] Z. Ferhat Hamida, Fake news detection on social media documents, Ph.D. thesis, 2024.

[16] F. Abid, C. Li, M. Alam, Multi-source social media data sentiment analysis using bidirectional recurrent convolutional neural networks, Computer Communications 157 (2020) 102–115.

[17] A. S. Gaafar, J. M. Dahr, A. K. Hamoud, Comparative analysis of performance of deep learning classification approach based on lstm-rnn for textual and image datasets, Informatica 46 (2022).

[18] G. Meena, K. K. Mohbey, A. Indian, Categorizing sentiment polarities in social networks data using convolutional neural network, SN Computer Science 3 (2022) 116.

[19] O. Edo-Osagie, I. Lake, O. Edeghere, B. De La Iglesia, Attention-based recurrent neural networks (rnns) for short text classification: An application in public health monitoring, in: Advances in Computational Intelligence: 15th International Work-Conference on Artificial Neural Networks, IWANN 2019, Gran Canaria, Spain, June 12-14, 2019, Proceedings, Part I 15, Springer, 2019, pp. 895–911.

[20] M. Ebrahimi, C. Y. Suen, O. Ormandjieva, Detecting predatory conversations in social media by deep convolutional neural networks, Digital Investigation 18 (2016) 33–49.

[21] A. Sallah, S. Agoujil, M. A. Wani, M. Hammad, A. A. Abd El-Latif, Y. Maleh, et al., Fine-tuned understanding: Enhancing social bot detection with transformer-based classification, IEEE Access (2024).

[22] A. K. Chowdhury, S. R. Sujon, M. S. S. Shafi, T. Ahmmad, S. Ahmed, K. M. Hasib, F. M. Shah, Harnessing large language models over transformer models for detecting bengali depressive social media text: A comprehensive study, Natural Language Processing Journal 7 (2024) 100075.

[23] P. Bhandari, Advancing intrinsic and non-intrinsic bug classification with NLP, machine learning, and few-shot prompt engineering, Ph.D. thesis, University of British Columbia, 2024.

[24] L. Yan, Y. Zheng, J. Cao, Few-shot learning for short text classification, Multimedia Tools and Applications 77 (2018) 29799–29810.

[25] M. Yu, X. Guo, J. Yi, S. Chang, S. Potdar, Y. Cheng, G. Tesauro, H. Wang, B. Zhou, Diverse few-shot text classification with multiple metrics, arXiv preprint arXiv:1805.07513 (2018).

[26] Y. Bao, M. Wu, S. Chang, R. Barzilay, Few-shot text classification with distributional signatures, arXiv preprint arXiv:1908.06039 (2019).

[27] O. Gramopadhye, S. S. Nachane, P. Chanda, G. Ramakrishnan, K. S. Jadhav, Y. Nandwani, D. Raghu, S. Joshi, Few shot chain-of-thought driven reasoning to prompt llms for open ended medical

question answering, arXiv preprint arXiv:2403.04890 (2024).

[28] Z. Li, S. Fan, Y. Gu, X. Li, Z. Duan, B. Dong, N. Liu, J. Wang, Flexkbqa: A flexible llm-powered framework for few-shot knowledge base question answering, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 2024, pp. 18608–18616.

[29] L. Zhou, S. Pan, J. Wang, A. V. Vasilakos, Machine learning on big data: Opportunities and challenges, Neurocomputing 237 (2017) 350–361.

[30] R. K. Mahabadi, L. Zettlemoyer, J. Henderson, M. Saeidi, L. Mathias, V. Stoyanov, M. Yazdani, Perfect: Prompt-free and efficient few-shot learning with language models, arXiv preprint arXiv:2204.01172 (2022).

[31] E. Loginova, W. K. Tsang, G. van Heijningen, L.-P. Kerkhove, D. F. Benoit, Forecasting directional bitcoin price returns using aspect-based sentiment analysis on online text data, Machine Learning 113 (2024) 4761–4784.

[32] D. Maynard, K. Bontcheva, D. Rout, Challenges in developing opinion mining tools for social media, Proceedings of@ NLP can u tag# usergeneratedcontent (2012) 15–22.

[33] E. Jiang, K. Olson, E. Toh, A. Molina, A. Donsbach, M. Terry, C. J. Cai, Promptmaker: Prompt-based prototyping with large language models, in: CHI Conference on Human Factors in Computing Systems Extended Abstracts, 2022, pp. 1–8.

[34] H. Qin, Revolutionizing cryptocurrency operations: The role of domain-specific large language models (llms), International Journal of Computer Trends and Technology 72 (2024) 101–113.

[35] S. J. Yoo, S. Lee, Large language models show human-like abstract thinking patterns: A construal-level perspective, in: Proceedings of the Annual Meeting of the Cognitive Science Society, volume 46, 2023.

[36] T. R. McIntosh, T. Liu, T. Susnjak, P. Watters, M. N. Halgamuge, A reasoning and value alignment test to assess advanced gpt reasoning, ACM Transactions on Interactive Intelligent Systems (2024).