Overview of CryptOQA: Opinion Extraction and Question Answering from CryptoCurrency-Related Tweets, Reddit, and Youtube Posts

Dhruv Kumar[1], Somrupa Sarkar[1], Koustav Rudra[1,*] and Kripabandhu Ghosh[2]

[1]IIT Kharagpur, Kharagpur, India
[2]Computational and Data Sciences (CDS), IISER Kolkata, Mohanpur, West Bengal, India

Abstract

Cryptocurrency continues to dominate online discourse, with platforms such as Twitter, Reddit, and YouTube serving as major hubs for public opinion, market speculation, and user-driven Q&A. The CryptoQA track at FIRE 2025 aims to develop systems that can automatically analyze and categorize cryptocurrency-related social media posts. This year, the track consisted of two tasks: (a) classifying posts into a three-level hierarchical label space covering categories such as Noise, Objective, Subjective (and its subtypes), and (b) detecting whether an answer is relevant to a given question in question–answer pairs. In total, five teams participated in the track, submitting a wide range of approaches which were evaluated primarily using the macro F1-score across all datasets and task levels. The participating systems demonstrated strong performance in handling challenges such as sentiment ambiguity, noisy and unstructured text, and multi-platform variability, offering valuable insights into modeling cryptocurrency discourse on social media.

Keywords

Cryptocurrency, Information Retrieval, Classification, Question Answering, Social Media

1. Introduction

Over the past decade, the emergence of new cryptocurrencies has had a profound impact on digital financial ecosystems. This shift has sparked ongoing conversations across various social media platforms. These platforms, such as Twitter, Reddit, and YouTube, are key places where users share opinions, ask questions, and discuss updates related to cryptocurrency technologies and market changes. The amount, variety, and speed of user-generated content create a significant and challenging issue for researchers who want to analyze large data streams consisting of text, images, and videos. Previous studies show that social media discussions about cryptocurrencies often reflect a mix of feelings, ranging from positive and negative responses to neutral comments, factual statements, ads, and user questions [1, 2].

Classifying these sentiments can offer useful insights into public opinion, new market trends, and user behavior. Correctly identifying subjective and objective content has been found to aid decision-making, improve market analysis, and allow better tracking of cryptocurrency-related conversations [3]. However, sentiment classification in this area remains challenging due to the unstructured nature of social media text. Posts often feature abbreviations, slang, evolving technical terms, and stylistic differences that make language interpretation difficult. Traditional text classification models often fail to capture these subtleties, especially when the expressions are short, casual, or dependent on context [4].

In addition to opinionated content, cryptocurrency discussions also include question-and-answer exchanges, where users seek guidance or clarification about investments, market conditions, or technical details. Often, community responses can be incomplete, irrelevant, or misleading. Thus, determining

Forum for Information Retrieval Evaluation, December 17-20, 2025, Varanasi, India
*Corresponding author.
pvt.dhruvkumar@gmail.com (D. Kumar); somrupas018@gmail.com (S. Sarkar); krudra@ai.iitkgp.ac.in (K. Rudra); kripaghosh@iiserkol.ac.in (K. Ghosh)
0009-0006-3989-5624 (D. Kumar)

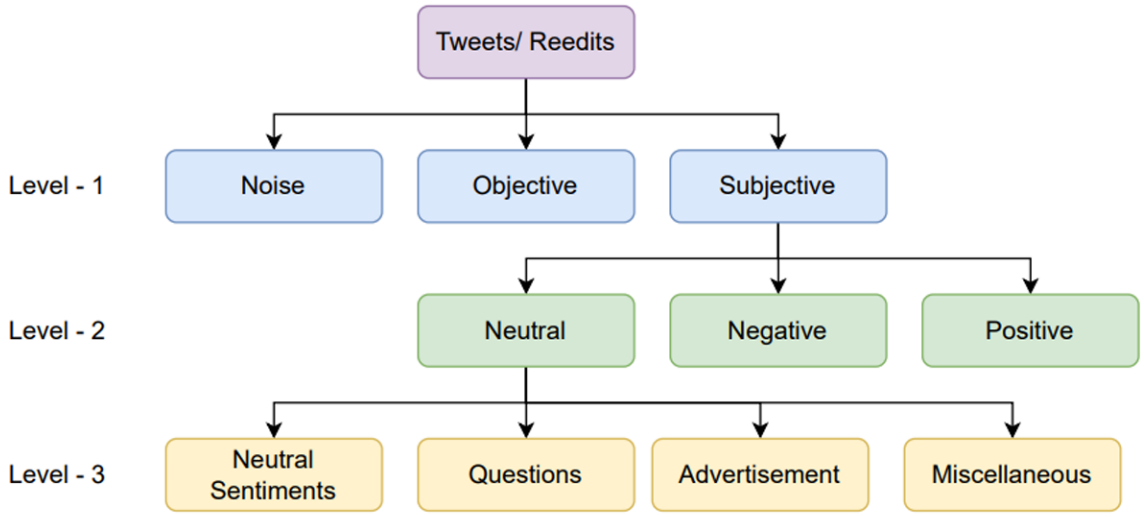

Figure 1: The Hierarchical Labeling of Opinions

whether a response effectively addresses the posed question is crucial for ensuring reliable access to information in cryptocurrency communities [5].

To tackle these issues, the aim of the proposed research is focused on creating systems that can perform two main tasks: (i) detailed classification of cryptocurrency-related posts into a hierarchical set of labels that differentiate noise, objective information, and subjective opinions, and (ii) assessing the relevance of responses in question-and-answer pairs. These tasks aim to support automated and ongoing monitoring of social media discussions, fostering a more organized understanding of cryptocurrency narratives across various platforms.

2. Dataset

There are three sources of datasets, namely Reddit, Twitter, and YouTube, which contain social media posts related to cryptocurrency. Now this dataset is again divided for classification and Q&A tasks, respectively. The classification dataset has three-level annotations:

1. **Level 1:** In level 1, there are three classes: Noise, Objective, and Subjective, and these three classes are marked with 0, 1, and 2, respectively.
2. **Level 2:** In this level, the Subjective class is further divided into three categories: Neutral, Negative, and Positive. These are marked with 0, 1, and 2, respectively, in the dataset.
3. **Level 3:** In the last level, there are four classes: Neutral Sentiments, Questions, Advertisement, and Miscellaneous, and these are marked with 0, 1, 2, and 3, respectively. This set of classes is branched from the Neutral category in level 2.

The hierarchical data distribution in the Twitter, Reddit, and YouTube datasets is shown in Figure 1. The Q&A task has a total of 31,692 samples across all the data sources (Twitter, Reddit, and YouTube combined). This dataset is further classified as Relevant or non-relevant, with 25,369 and 6,323 samples from the training and test sets, respectively. The dataset contains 3,704 datapoints for the Relevant class and 21,531 datapoints for the Non-relevant class.

2.1. Training data statistics

The training data provided for this evaluation contain posts from Reddit, Twitter, and YouTube, annotated across three hierarchical levels and mapped to eight final categories. The distribution for each platform is reported in Figures 2, 3, and 4.

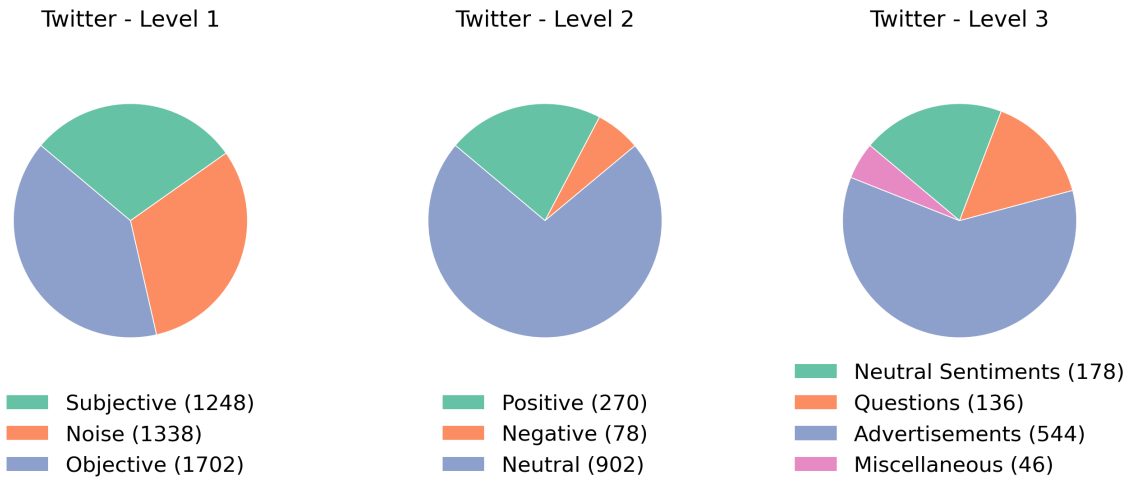

Figure 2: Twitter dataset distribution in three levels

For Reddit, the Level 1 labels contain 645 Noise, 503 Objective, and 3,852 Subjective posts. In Level 2, the Subjective class is divided into 259 Positive, 410 Negative, and 3,183 Neutral posts. In Level 3, the Neutral posts are further split into 476 Neutral Sentiments, 2390 Questions, 105 Advertisements, and 212 Miscellaneous samples.

For Twitter, the Level 1 distribution includes 1,338 Noise, 1,702 Objective, and 1,248 Subjective posts. These Subjective posts are divided in Level 2 into 270 Positive, 78 Negative, and 902 Neutral posts. The Level 3 breakdown of Neutral posts consists of 178 Neutral Sentiments, 136 Questions, 544 Advertisements, and 46 Miscellaneous entries.

For YouTube, Level 1 consists of 786 Noise, 32 Objective, and 4,182 Subjective posts. In Level 2, the Subjective posts include 207 Positive, 1,574 Negative, and 2,401 Neutral samples. Level 3 further categorizes the Neutral posts into 1,391 Neutral Sentiments, 1,000 Questions, 1 Advertisement, and 9 Miscellaneous posts.

2.2. Test data statistics

The test split contains 500 posts each from Reddit, Twitter, and YouTube, totaling 1,500 posts across platforms. Each platform follows the same three-level hierarchical label design as the training data, ensuring consistency for evaluation.

For Q&A, the dataset includes 6,323 paired entries consisting of a question, its corresponding comment, and a binary relevance label indicating whether the comment addresses the question.

3. Task Definition

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, and Miscellaneous.

Task 2 required participants to identify all answers relevant to a given question on cryptocurrency.

4. Participants

There are final submissions from five teams representing various academic institutions in the CryptOQA shared task at FIRE 2025, focusing on classifying social media posts related to cryptocurrency. The varied strategies and advanced mathematical models employed by the respective teams to deal with the given task are mentioned below:

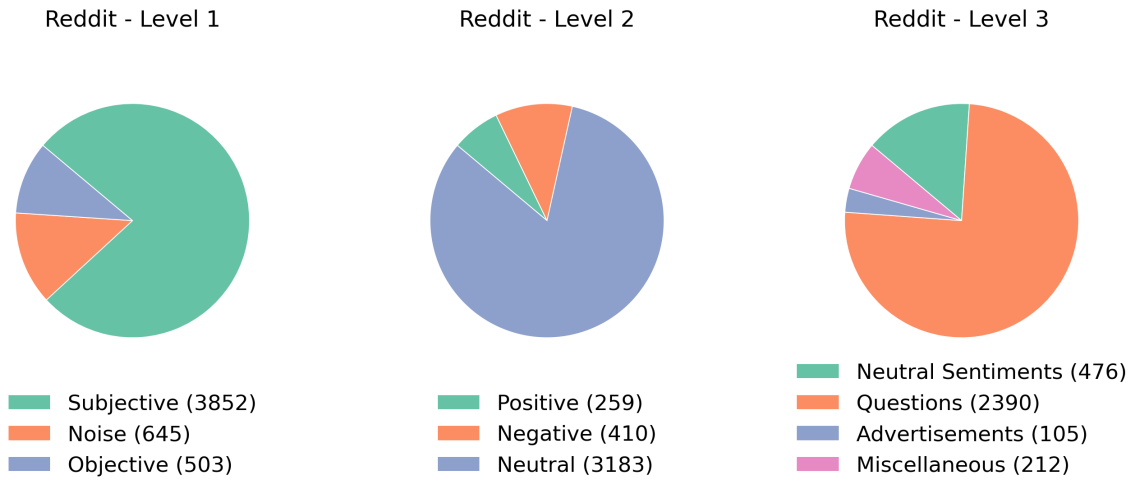

Figure 3: Reddit dataset distribution in three levels

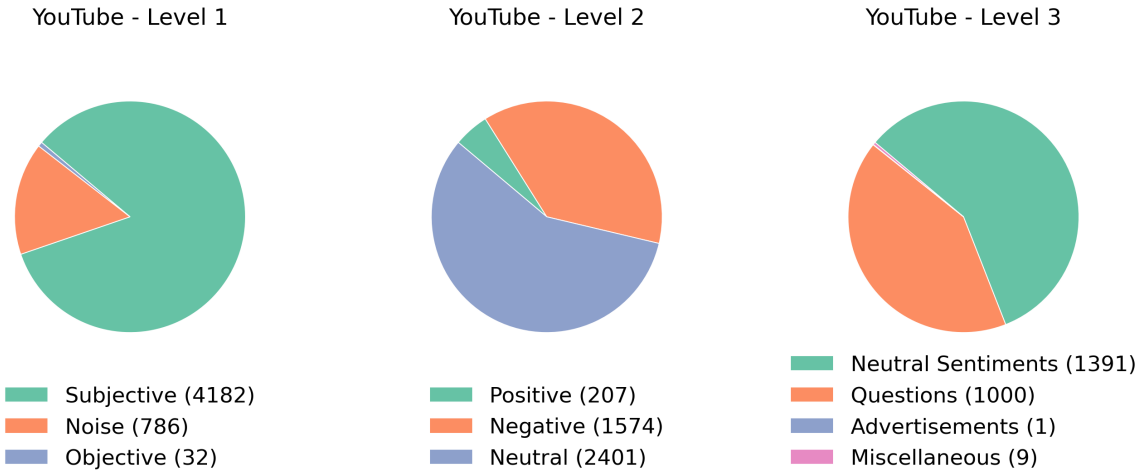

Figure 4: YouTube dataset distribution in three levels

1. **Team KLU** (Koneru Lakshmaiah University) submitted a multi-stage transformer architecture built on DeBERTa-V3-Small [6], where a shared encoder feeds multiple task-specific heads for each hierarchical level as well as for relevance prediction. Their design integrates focal loss [7], Dice loss [8], supervised contrastive learning [9], and label smoothing [10] within a unified multi-loss framework, and uses conditional routing between levels. The system also employs 5-fold ensembling and layer-wise learning-rate decay for stable optimization. This approach achieved the highest performance in the shared task, ranking first with a macro F1 of 1.00 in Task 1 and 1.00 in Task 2.
2. **Team NLPFusion** (Mangalore University) implemented a multi-task transformer model using BERT [11] and CryptoBERT encoders to jointly learn all three hierarchical levels, supported by task-specific recurrent modules and shared optimization across levels. Their relevance prediction system uses transformer encoders to process concatenated question–answer pairs. This joint modeling strategy resulted in competitive performance, with the team ranking second in both tasks, with a macro F1 of 0.6253 in Task 1 and 0.7940 in Task 2.
3. **Team IISER** (Indian Institute of Science Education and Research) adapted the Gemma-1B LLM [12] for 8-class flat classification by replacing the original output projection layer with a custom classification head and fine-tuning only the final transformer block and normalization layers. Their submission includes an extensive data-cleaning pipeline covering structural corrections, noise removal, and text normalization to improve robustness. A weighted cross-entropy loss

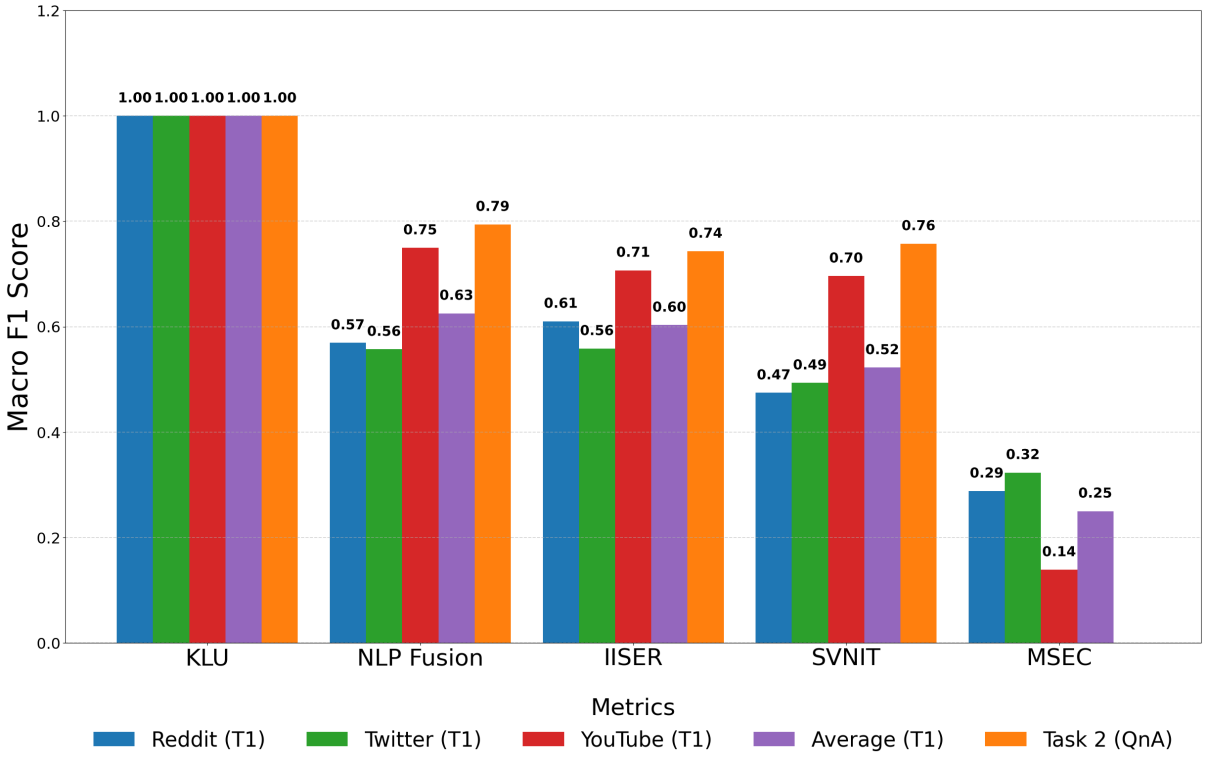

Figure 5: Team Performance Across All Tasks, here T1 refers to Task 1

function was used to mitigate class imbalance. This efficient fine-tuning strategy produced a macro F1 of 0.6028, securing the third rank in Task 1 and a macro F1 of 0.7432, securing the fourth rank in Task 2.

4. **Team SVNIT** (Sardar Vallabhbhai National Institute of Technology) explored recurrent neural architectures with Word2Vec [13], GloVe [14], and FastText [15] embeddings for hierarchical classification, ultimately selecting a GRU-based model [16] enhanced with attention [17] for Level-wise predictions. For the Q&A task, they employed a Siamese GRU network [18] that embeds questions and answers and predicts their semantic relevance. Their recurrent approach performed reliably on both subtasks, yielding a Task 1 macro F1 of 0.5222, securing the fourth rank, and a Task 2 macro F1 of 0.7575, securing the third rank in this subtask.
5. **Team MSEC** (Meenakshi Sundararajan Engineering College) used a BERT-base (uncased) model [11] fine-tuned as a single recursive classifier predicting all hierarchical labels within a unified output layer. Their approach incorporates lowercasing, WordPiece tokenization [19], attention masks, AdamW optimization [20], and dropout regularization to adapt the pretrained model to cryptocurrency-related social media text. The system obtained a Task 1 macro F1 of 0.2500, while no submission was made for Task 2.

5. Methodologies

The submitted solutions across teams participating in the CryptOQA shared task employed a range of techniques for classifying cryptocurrency-related social media posts and determining the relevance of question–answer pairs. These methodologies can be broadly categorized into four techniques, namely, transformer-based models, hierarchical classification, recurrent neural architectures (LSTM/GRU), and parameter-efficient LLM fine-tuning. To tackle the challenges posed by noisy, multi-platform data and the multi-level label structure, each team adopted one or more of these methodological strategies.

Transformer-based Models Transformer architectures remained the dominant choice due to

their ability to capture long-range dependencies and contextual relationships through self-attention. These models are particularly well-suited for domain-specific and informal language prevalent in cryptocurrency discourse. Several teams fine-tuned pre-trained transformer encoders, such as DeBERTa and BERT, for both subtasks.

- **DeBERTa-V3-Small** was used by Team KLU, integrating multiple task-specific classification heads and specialized loss functions to improve discriminability across the hierarchical label space.
- **BERT and CryptoBERT** were employed by Team NLPFusion in a multi-task learning framework, allowing the model to jointly learn representations for all hierarchical levels and Q&A relevance prediction.
- **BERT-base** (uncased) was utilized by Team MSEC as the primary encoder for a single unified hierarchical classifier fine-tuned on cleaned and normalized text from all platforms.

Hierarchical Classification Several teams incorporated the hierarchical structure of the CryptOQA labels directly into their modeling pipelines. Hierarchical classification enables predictions to be made at increasing levels of granularity while reducing confusion between semantically distant classes.

- **Hierarchical multi-head design** was implemented by Team KLU, in which Level 1 outputs are conditionally routed to deeper classification heads at Levels 2 and 3, enabling fine-grained discrimination among sentiment and question categories.
- **Multi-task hierarchical modeling** was adopted by Team NLPFusion, where shared transformer encoders simultaneously optimize for Level 1, Level 2, Level 3, and Q&A loss functions, leveraging inter-level dependencies without requiring separate models for each stage.
- **Recursive hierarchical prediction** was incorporated in Team MSEC's BERT-based system, where all hierarchical outputs are produced within a single unified classifier, simplifying the inference pipeline.

Recurrent Neural Models (LSTM/GRU) Recurrent neural networks were also explored due to their effectiveness in modeling sequential dependencies and sentiment-bearing patterns in short social media posts. These models remain competitive for noisy and domain-specific text when combined with appropriate embeddings.

- **GRU-based hierarchical classifiers** were adopted by Team SVNIT, enhanced with attention mechanisms to focus on discriminative tokens across levels.
- **Siamese GRU networks** were applied by the same team for the Q&A task, embedding questions and answers separately and computing their semantic alignment for relevance prediction.
- **LSTM-based architectures** were explored in initial experimentation across some teams, although GRU variants generally provided more stable performance in the final submitted runs.

Parameter-efficient LLM Fine-tuning With the availability of compact open-weight LLMs, teams also explored parameter-efficient tuning approaches to reduce computational cost while retaining strong contextual understanding.

- **Gemma-1B** was adapted by Team IISER by replacing the model's output projection with an eight-class head and updating only the final transformer block along with normalization layers, significantly reducing training overhead.
- **Selective layer unfreezing and lightweight optimization strategies** were employed to stabilize fine-tuning on the small task-specific datasets while mitigating overfitting.

These approaches were accompanied by extensive data preprocessing steps, including text cleaning, noise filtering, token normalization, and class-weighted loss functions, to compensate for the limited parameter updates.

Team Name	Macro F1 Score				
	Task 1				Task 2
	Reddit	Twitter	YouTube	Avg.	
KLU	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
NLP Fusion	0.5694	0.5570	0.7496	0.6254	0.7940
IISER	0.6102	0.5584	0.7070	0.6029	0.7432
SVNIT	0.4744	0.4933	0.6961	0.5222	0.7575
MSEC	0.2881	0.3232	0.1388	0.2500	–

Table 1
Best reported results from the respective teams. Here, task 1 refers to the classification task and task 2 refers to Q&A task. Scores in bold are the best performers in this track.

6. Results

The results from the CryptOQA 2025 shared task highlight that transformer-based models demonstrate clear superiority in categorizing cryptocurrency-related social media posts. Team KLU achieved the best performance in Task 1, attaining a perfect macro F1 score of 1.00 across Reddit, Twitter, and YouTube using a multi-stage DeBERTa-V3–based hierarchical architecture enriched with multiple loss functions and ensembling. Their system substantially outperformed the remaining submissions. Team NLPFusion followed with a macro F1 of 0.6253, supported by a multi-task transformer framework using BERT and CryptoBERT encoders, while Team IISER's parameter-efficient adaptation of Gemma-1B secured a close macro F1 of 0.6028. Recurrent neural models performed moderately well, with Team SVNIT reporting a macro F1 score of 0.5222, indicating that GRU-based architectures can still be competitive on noisier platforms, despite lacking the representational power of transformers. Team MSEC's unified BERT-based classifier achieved a macro F1 of 0.2500, indicating difficulty in handling the deeper sentiment granularity required at Level 2 and Level 3. Overall, transformer-driven approaches remained the most successful for hierarchical classification.

For the Question and Answering scenario in Task 2, the findings exhibit clearer variation across architectures. Team KLU again attained the highest score with a macro F1 of 1.00, demonstrating the strength of supervised contrastive and multi-loss training strategies for semantic relevance prediction. Team NLPFusion achieved the next best performance with 0.7940, followed by Team SVNIT's Siamese GRU network, which obtained 0.7575, reflecting that recurrent similarity models remain effective for paired-sentence inference. Team IISER's lightweight Gemma-1B fine-tuned system reported 0.7432, showing that selective parameter tuning can still yield strong generalization. Team MSEC did not submit a ranked run for Task 2. In general, transformer-based systems dominated Task 1 with significantly higher scores, while both transformers and recurrent Siamese models showed competitive performance in the Q&A task. All contributions and their final results are listed in Table 1.

7. Conclusion

The CryptOQA 2025 shared task aimed to evaluate machine learning and NLP approaches for hierarchical post classification and Q&A relevance detection in cryptocurrency-related social media content. Transformer-based models have proven to be most effective, with architectures built on DeBERTa-V3, BERT, and CryptoBERT delivering the strongest results, exemplified by Team KLU's multi-stage transformer system achieving perfect macro F1 scores in both tasks. While recurrent architectures, such as GRU-based classifiers, performed reliably, they remained less effective than transformers for making deeper sentiment distinctions. Parameter-efficient LLM tuning, as seen with Gemma-1B, also showed competitive potential, particularly under computational constraints. Hierarchical modeling benefited from multi-task transformer strategies, though label imbalance and fine-grained categories continued to pose challenges. For Q&A relevance, both transformer models and Siamese GRU networks produced strong outcomes, with transformers ultimately leading.

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly in order to: Grammar and spelling check, and reword. 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] D. Kim, Effects of Social Media-Based Peer Opinions on the Prices of Cryptocurrency Options, Journal of Futures Markets 45 (2025) 1512–1543.

[2] S. Rouhani, E. Abedin, Crypto-Currencies Narrated on Tweets: A Sentiment Analysis Approach, International Journal of Ethics and Systems 36 (2020) 58–72.

[3] S. Yu, J. Padfield, Advanced Techniques in Profiling Cryptocurrency Influencers: A Review, International Journal of Blockchains and Cryptocurrencies 5 (2024) 1–18.

[4] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks, arXiv:1908.10084 (2019).

[5] X. Hu, Enhancing Answer Selection in Community Question Answering with Pre-Trained and Large Language Models, arXiv:2311.17502 (2023).

[6] P. He, X. Liu, J. Gao, W. Chen, DEBERTA: Decoding-Enhanced BERT with Disentangled Attention, in: International Conference on Learning Representations, 2020.

[7] T. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal Loss for Dense Object Detection, IEEE Transactions on Pattern Analysis & Machine Intelligence 42 (2020) 318–327.

[8] F. Milletari, N. Navab, S. Ahmadi, V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation, in: 4th International Conference on 3D Vision (3DV), 2016, pp. 565–571.

[9] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised Contrastive Learning, Advances in Neural Information Processing Systems 33 (2020) 18661–18673.

[10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the Inception Architecture for Computer Vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.

[11] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 2019, pp. 4171–4186.

[12] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Riviere, M. S. Kale, J. Love, et al., Gemma: Open Models Based on Gemini Research and Technology, arXiv:2403.08295 (2024).

[13] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, arXiv:1301.3781 (2013).

[14] J. Pennington, R. Socher, C. D. Manning, GloVe: Global Vectors for Word Representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.

[15] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching Word Vectors with Subword Information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146.

[16] K. Cho, B. Van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation, arXiv:1406.1078 (2014).

[17] D. Bahdanau, K. Cho, Y. Bengio, Neural Machine Translation by Jointly Learning to Align and Translate, arXiv:1409.0473 (2014).

[18] J. Mueller, A. Thyagarajan, Siamese Recurrent Architectures for Learning Sentence Similarity, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 30, 2016.

[19] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al., Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation, arXiv:1609.08144 (2016).
[20] I. Loshchilov, F. Hutter, Decoupled Weight Decay Regularization, arXiv:1711.05101 (2017).