Overview of CryptOQA: Opinion Extraction and Question Answering from CryptoCurrency-Related Tweets and Reddit posts

Gourav Sen[1,†], Sougata Sarkar[2,†], Subham Kumar[3], Koustav Rudra[3,*] and Kripabandhu Ghosh[5]

[1]IIT Jodhpur, Jodhpur, Rajasthan, India
[2]Deloitte, Kolkata, West Bengal, India
[3]IIT Kharagpur, Kharagpur, West Bengal, India
[5]IISER Kolkata, Mohanpur, West Bengal, India

Abstract

Cryptocurrency is an exponentially growing domain, and Twitter and Reddit are important outlets for users to express their opinions, ask questions, or discuss specific topics. The CryptoQA track at FIRE 2024 aims to develop systems that automatically assess cryptocurrency posts on social media. In this track, there are basically two tasks: (a) the classification of posts related to cryptocurrencies into the eight classes, namely, Noise, Objective, Positive, Negative, Neutral, Question, Advertisement, and Miscellaneous, divided across three levels, and (b) the participants had to detect whether the answer is relevant to the question given question-answer pair.

Overall, three teams from various institutions have submitted to this track. These experiments were evaluated using several quantitative measures, particularly the F1-Score, as well as a qualitative evaluation based on features developed, supervised learning models, and their learning parameters. These models were able to successfully tackle the issues of sentiment dispersion, noise, and multiple versions of posts, which significantly helped in mitigating the lingering noise from the crowd in the context of cryptocurrency.

Keywords

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

1. Introduction

Over the last decade, the emergence of new cryptocurrencies has significantly altered the evolution of the global economy, triggering complex debates on multiple online platforms. Social networks [1] like Reddit and Twitter are one of the primary sources for public discourse on cryptocurrency market changes, trends, and technologies, offering a valuable yet challenging task to analyze big data streams in the form of text, images, videos, etc, for researchers. A broad spectrum of sentiments [2] is posed on the social media posts related to cryptocurrencies, and these span over various formats such as questions, enquiries, opinions, advertisements, or objective statements [3].

By classifying the sentiments of these posts, individuals and organizations can offer better insights, offering insightful trends about public opinion. Accurate classification of sentiments ranging from positive and negative to neutral leads to predicting market trends, understanding consumer behaviour, and shaping marketing strategies [4] effectively. However, the sentiment classification task in this context attracts many challenges due to its diverse and unstructured nature of social media content. One of the primary challenges is the inherent variability and ambiguity of social media language, which often involves abbreviations and slang words. It is difficult for a traditional text classification model to accurately interpret the nuances of expression in the social media posts that can be short, informal and slang-laden [5]. Furthermore, the task gets complicated as the sentiment in which posts can be expressed

Forum for Information Retrieval Evaluation, December 12-15, 2024, India
*Corresponding author.
†These authors contributed equally.
sengourav0704@gmail.com (G. Sen); sougata.sarkar8101@gmail.com (S. Sarkar); kumarshubham209@gmail.com (S. Kumar); krudra@ai.iitkgp.ac.in (K. Rudra); kripaghosh@iiserkol.ac.in (K. Ghosh)

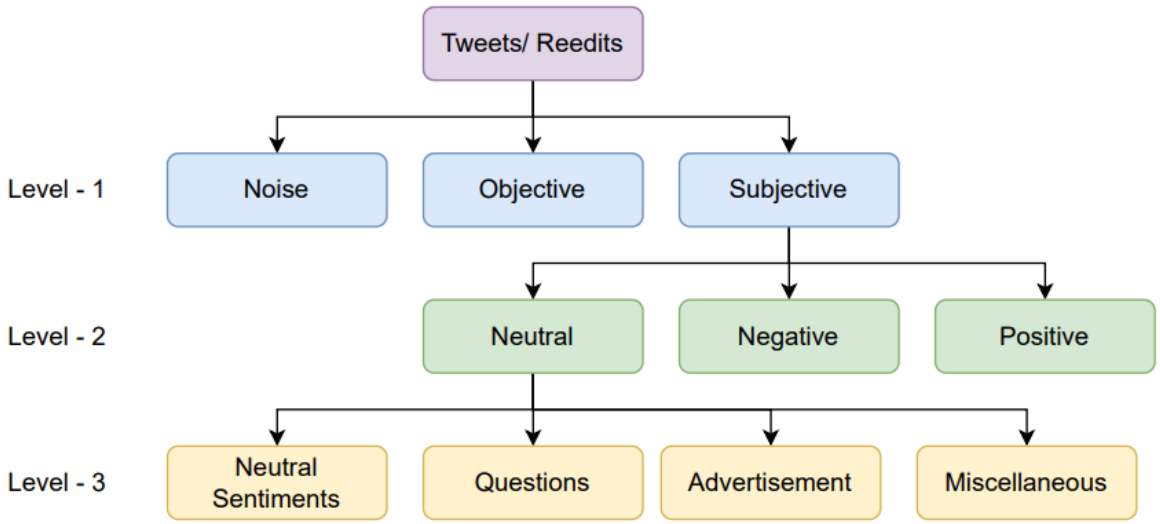


Figure 1: The Hierarchical Labeling of Opinions

differs significantly. These problems need to be addressed with advanced models [6] to account for the nuances of linguistics used in social media postings, thereby providing a refined classification of sentiments. Therefore, the aim of the proposed research is to develop a system that provides effective and precise automated monitoring of social media cryptocurrency discussions. The system that was developed should be able to classify the content in the opinion with fine-grained sentiment, fact, and opinion, as well as within the fact and opinion noise information. Moreover, the system must also be able to provide effective replies to questions related to cryptocurrency. In addition to the classification task, this track focused on addressing other queries and concerns that potential crypto investors may have, and attempted to determine whether the comment regarding the question was appropriate or not.

2. Dataset

There are two sources of datasets, namely, Reddit and Twitter social media posts. Now this dataset is again divided for classification and QnA tasks respectively. The classification dataset has three level-annotations:

1. **Level 1:** In level 1, there are three classes: NOISE, OBJECTIVE, SUBJECTIVE, and these three classes are marked with 0,1,2, respectively.
2. **Level 2:** In this level, the SUBJECTIVE class is further divided into three categories: NEUTRAL, NEGATIVE, POSITIVE and these are marked with 0,1,2, respectively, in the dataset.
3. **Level 3:** In the last level, there are four classes, namely, NEUTRAL SENTIMENTS, QUESTIONS, ADVERTISEMENTS, MISCELLANEOUS, and these are marked with 0,1,2,3, respectively. This set of classes is branched from the NEUTRAL category in level 2.

The hierarchical data distribution in Twitter and Reddit datasets is shown in Figure 1.

The QnA task has a total of 31,614 samples across both data sources (Twitter and Reddit combined). This dataset is further classified as Relevant or non-relevant, with 25,290 and 6,324 samples from the training and test sets, respectively.

2.1. Training data statistics

The distribution of training data for both Twitter and Reddit samples is labeled in three different levels among 8 diverse categories as depicted in Figure 2 and 3, respectively. For Twitter training data, there

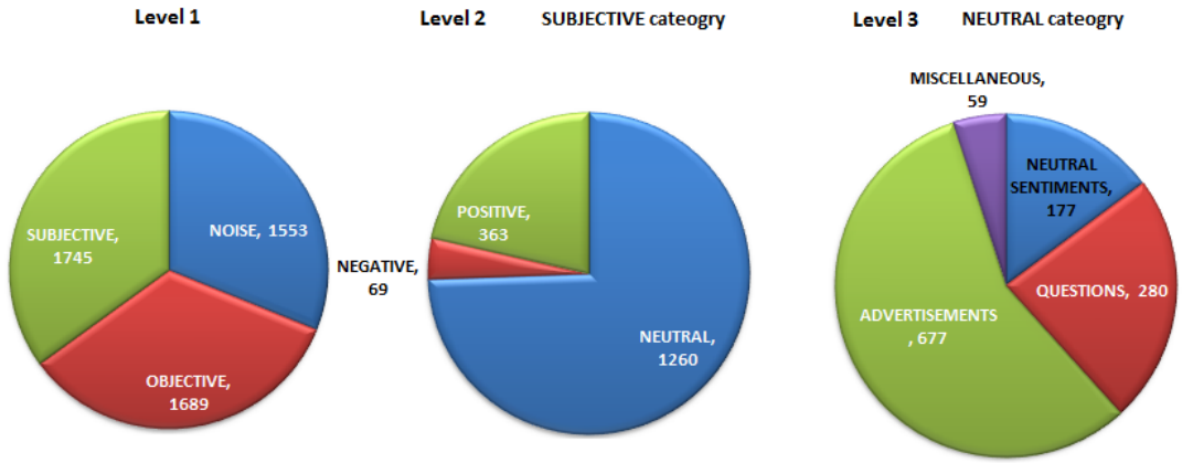


Figure 2: Twitter dataset distribution in three levels

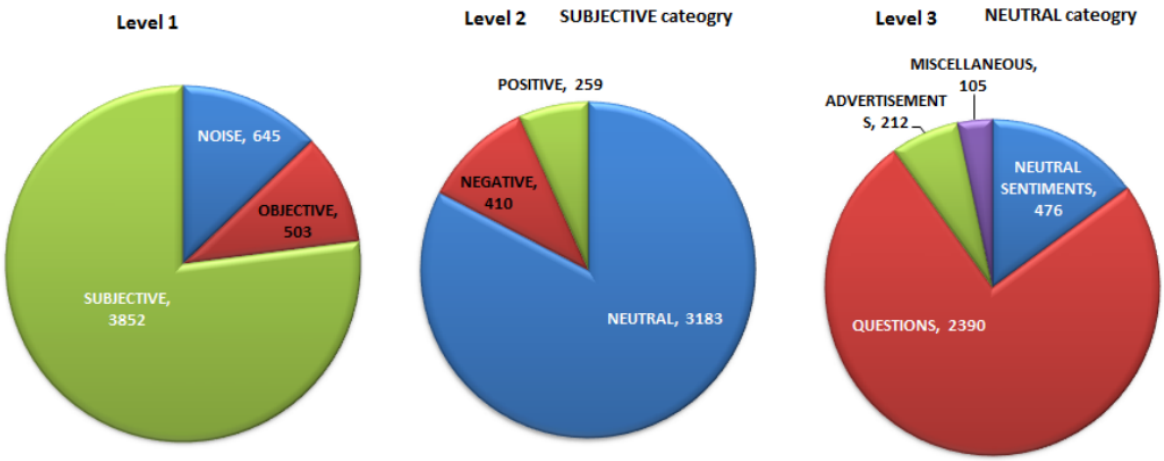


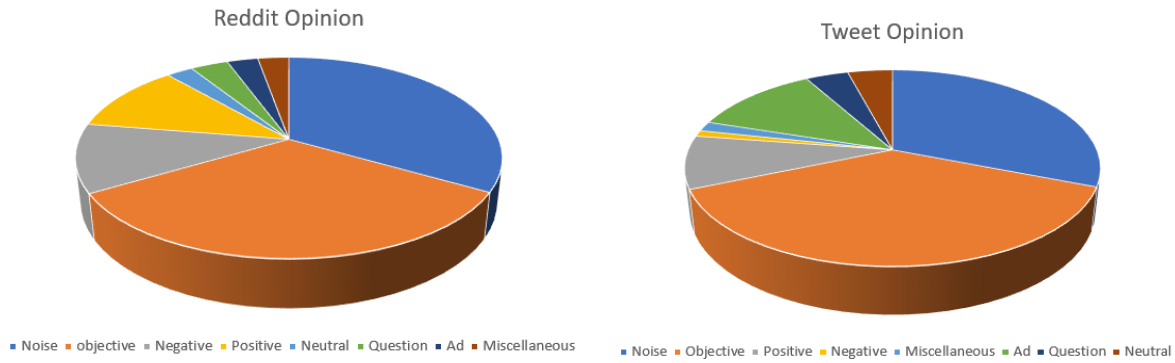
Figure 3: Reddit dataset distribution in three levels

are 1,745, 1,553, and 1,689 training samples of the SUBJECTIVE, NOISE, and OBJECTIVE classes in level 1. Furthermore, this subjective class is divided into three levels, with samples of 363, 69, and 1260 for the positive, negative, and neutral classes, respectively. In level 3, this NEUTRAL category is further classified as 177, 280, 677, and 59 training samples of NEUTRAL SENTIMENTS, QUESTIONS, ADVERTISEMENTS, and MISCELLANEOUS classes. Similarly, for Reddit, the distribution of samples across the NOISE, OBJECTIVE, POSITIVE, NEGATIVE, NEUTRAL, QUESTIONS, ADVERTISEMENTS, and MISCELLANEOUS classes is 645, 503, 259, 410, 476, 2,390, 212, and 105, respectively.

2.2. Test data statistics

The opinion test dataset contains two sub-datasets for each of the social media platforms (Twitter and Reddit). The total volume of the test dataset is 1000 texts (500 for each media).

QnA Dataset: The question-answering (QnA) dataset contains a set of question-answer pairs with labels marked as Relevant and Not Relevant for the answer being relevant to the question or not, respectively. The Question labels inspire the QnA corpus in the opinion dataset. The test dataset contains 6,324 QnA pairs, with 888 pairs considered relevant and 5,436 as non-relevant. Figure 4 shows the distribution of Reddit and Twitter data.



(a) **Reddit** (b) **Twitter**

Figure 4: Test data distribution for opinion classification task over Reddit and Twitter

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*, *Miscellaneous*.

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

4. Participants

There are final submissions of four teams from various academic institutions in the CryptOQA shared task at FIRE 2024, focused 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:

1. **Team MUCS** (Mangalore University) approaches this challenge with two deep learning models. (a) Unique_Label_LSTM, which has a unique labeling method for hierarchical classification, and (b) HCC_LSTM, a hierarchical classifier chain model which uses LSTM [7] internally. The latter model has achieved better performance over the former, reporting macro F1 scores of 0.574 and 0.328 for Twitter and Reddit, respectively.

2. **Team TextTitans** (IIT Kharagpur) leverages large language models (LLMs), namely, GPT-4-Turbo, for opinion classification and question-answering tasks. A 64-shot prompting technique is utilized to categorize social media posts [8]. They reported scores of 0.266 and 0.249 on Twitter and Reddit, respectively. However, they ranked 1st in the question-answering task with a score of 0.157.

3. **Team COM** presented two frameworks with transformer models namely, XLM-RoBERTa [9] for single-level classification and RoBERTa-base for 3-level hierarchical classification. Now, a RoBERTa is used in each level for the classification task. The posts were first classified as Noise, Objective, or Subjective at level 1. Subjective posts were further classified into Neutral, Negative, or Positive sentiments at level 2. Finally, Neutral posts were classified into Neutral-Sentiment, Questions, Advertisements, or Miscellaneous at level 3. In contrast, the single-level framework used an XLM-RoBERTa model to categorize posts into one of eight classifications. They obtained the scores of 0.778 and 0.542 on Twitter and Reddit opinion classification tasks, respectively.

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. These methodologies can be

broadly categorized into four techniques, namely, transformer-based models, hierarchical classification, LSTM models [10], and prompt-based learning. To address the challenges imposed, each team adopted one or more of these techniques.

Transformer-based Models These models are known for their capacity to capture context and relationships within text, leveraging attention mechanisms. This approach is a popular choice for most teams, as it effectively handles the nuanced and domain-specific language often found in cryptocurrency-related social media posts.

- *RoBERTa:* This model is heavily pre-trained on robust and large datasets to capture the intricacies of the text. RoBERTa is used by Team COM.
- *XLM-RoBERTa:* Team COM developed a multi-label classification framework for single levels through a fine-tuned XLM-RoBERTa-base model that employs a multilingual transformer model, which supersets RoBERTa in a hundred languages. This was done to leverage the ability of XLM-RoBERTa to generalize across a number of data formats.

Hierarchical Classification Hierarchical classification is the process of creating a tree-like structure for the classification problem that supports the construction of predictions at different granularities and handles the case of multi-label classification in a simple way. The teams that used this kind of approach partitioned the data into several levels, where each level processes more detailed differences.

- *Hierarchical Classifier Chain (HCC_LSTM):* This strategy was used by Team MUCS, an LSTM classifier at each level of the hierarchy to manage hierarchical relations.
- *3-level RoBERTa Hierarchical Framework:* Team COM has utilized this method to classify posts in growing depth of detail categories.

LSTM-based Models Long Short-Term Memory (LSTM) models have the ability to process sequential data and capture long-term dependencies in text with high relative distance, making them suitable for tasks where the context of the sentence is crucial.

- *Unique_Label_LSTM:* Team MUCS leverages this technique as a unique labeling technique for hierarchical classification.
- *BiLSTM for Question-Answering (QnA):* A BiLSTM model was used by Team COM to classify comments as relevant or non-relevant in the QnA task.

Prompt-based Learning and Few-shot Techniques Prompts were used in learning, especially in conjunction with large language models (LLMs) to utilize pre-trained knowledge. This approach is particularly beneficial in cases with no or a small amount of labeled data, as it directs the model in formulating answers in the form of a prompt. There are methods, namely zero-shot, where no data is provided, few-shot, where a limited number of samples are given and many more.

- *GPT-4-Turbo with 64-shot Learning:* Team TextTitans employed a few-shot learning technique [11] that allowed the model to classify with very few examples without requiring a significantly large labeled dataset.

6. Result

The results from the task CryptOQA highlight that transformer-based models are superior in categorizing social media posts related to cryptocurrency. Team COM had the best macro F1 scores for Task 1 where they attained 0.778 for Twitter and 0.542 for Reddit using an XLM-RoBERTa based single level classification framework.

Team MUCS used hierarchical classification models, which proved to be efficient as well, but Team COM's hierarchical model was only a fraction as successful as their single-level model. Lastly, the TextTitans team reported F1-scores of 0.266 and 0.249 on the Twitter and Reddit datasets, respectively.

Team Name	Tasks	Twitter F1-score	Reddit F1-score
COM	Task 1	0.778	0.542
	Task 2	0.146	0.146
MUCS	Task 1	0.574	0.328
	Task 2	-	-
TextTitans	Task 1	0.266	0.249
	Task 2	0.157	0.157

Table 1
Best reported results from the respective teams. Here, **Task 1** refers to the classification task and **Task 2** refers to the QnA task.

For the Question and Answering scenario in task 2, the findings are more varied. Team TextTitans reported the highest score of 0.157 when using a prompt-based approach on the GPT4-Turbo model. However, the performance of other teams, including Team COM's BiLSTM model, was notably lower, with a score of 0.146. In general, models based on transformers were the most successful in Task 1, achieving higher scores, while prompt-based techniques showed their dominance in QnA tasks. All the contributions are listed in Table 1.

7. Conclusion

The goal of the CryptOQA task is to evaluate the post classification and QnA tasks related to cryptocurrencies using ML and NLP techniques. Within the submissions, transformer-based approaches, such as RoBERTa and XLM-RoBERTa, consistently yield better results. The use of transformers yielded the highest scores across the Twitter and Reddit datasets, with Team COM topping the performance using RoBERTa for single-layer classification. Sensitive to those changes were the hierarchical classification models, which still performed poorly compared to their single-level counterparts. Approaches based on prompt learning methods, especially the few-shot models, worked well in the QnA task. Overall, these findings are important as a basis for further development in the rapidly evolving cryptocurrency market.

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] H. N. Durmuş Şenyapar, Cryptocurrency on social media: Analyzing the digital discourse towards the coin market 9 (2024) 202–223.

[2] 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). URL: https://www.mdpi.com/2504-2289/8/6/63. doi:10.3390/bdcc8060063.

[3] L. Nizzoli, LEVERAGING SOCIAL MEDIA AND AI TO FOSTER SECURE SOCIETIES AGAINST ONLINE AND OFFLINE THREATS, Ph.D. thesis, 2021. doi:10.13140/RG.2.2.29807.97446.

[4] S. Oikonomopoulos, K. Tzafilkou, D. Karapiperis, V. Verykios, Cryptocurrency price prediction using social media sentiment analysis, 2022, pp. 1–8. doi:10.1109/IISA56318.2022.9904351.

[5] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL: https://arxiv.org/abs/1908.10084. arXiv:1908.10084.

[6] G. Kim, D.-H. Shin, J. G. Choi, S. Lim, A deep learning-based cryptocurrency price prediction model that uses on-chain data, IEEE Access 10 (2022) 56232–56248. doi:10.1109/ACCESS.2022.3177888.

[7] N. Aslam, F. Rustam, E. Lee, P. B. Washington, I. Ashraf, Sentiment analysis and emotion detection on cryptocurrency related tweets using ensemble lstm-gru model, IEEE Access 10 (2022) 39313–39324. doi:10.1109/ACCESS.2022.3165621.

[8] 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 Comput. Surv. 55 (2023). URL: https://doi.org/10.1145/3560815. doi:10.1145/3560815.

[9] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, 2020. URL: https://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[10] X. Huang, W. Zhang, X. Tang, M. Zhang, J. Surbiryala, V. Iosifidis, Z. Liu, J. Zhang, Lstm based sentiment analysis for cryptocurrency prediction, 2021. URL: https://arxiv.org/abs/2103.14804. arXiv:2103.14804.

[11] 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, 2024. URL: https://arxiv.org/abs/2308.12060. arXiv:2308.12060.