Transformer based Opinion Classification of Cryptocurrency Posts

C.Jerin Mahibha[1,*], Wordson Robert[2], Gersome Shimi[3] and Durairaj Thenmozhi[4]

[1]Meenakshi Sundararajan Engineering College, Chennai
[2]Indian Institute of Science Education and Research, Kolkotta, India
[3]Madras Christian College,Chennai, India
[4]Sri Sivasubramaniya Nadar College of Engineering, Chennai

Abstract

Data classification is the process by which given phrases or sentences are contextualized and mapped into predefined, meaningful categories. This is commonly achieved through natural language processing techniques such as sentiment analysis, which interprets the emotional tone or intent behind textual content. In online spaces and chatrooms, where language tends to be informal and context-heavy, such a classification becomes essential for a range of applications, including moderation, content personalization, and automated policy enforcement. However, the presence of sarcasm, slang, and ungrammatical constructions in user-generated content significantly complicates this task. Posts often lack a clear structure, making it difficult for conventional models to extract reliable semantic cues. In response to this challenge, we propose a classification model fine-tuned on cryptocurrency-related discussions sourced from diverse social media platforms. Developed as part of the FIRE 2025 shared task organized by CryptOQ, our system employs a 12-layer English monolingual general-purpose transformer model — specifically, BERT-base (uncased) — pretrained on lowercased English text from Wikipedia and BookCorpus.

1. Introduction

Data classification, refers to the process where the data are grouped into the necessary categories. Multiple levels of classification are done to analyze the categorization further with much more clarity. The proposed system is associtaed with opinion classification on a crptocurrency related posts from social media, including Twitter, Reddit and YouTube.

The same model could be potentially used to classify online conversations about different topics. The model used by the proposed system had to be fine-tuned on online verbiage, highly stylized phrases, and even potentially misleading information. To accomplish this, a process known as sentiment analysis was employed, which essentially provides a contextual framework for previously unstructured data. This analysis enabled to decipher the underlying intent behind the data, empowering to effective utilization of it for various needs.

The model proposed for the FIRE 2025 shared task, organized by CryptOQ, uses a 12-layer English general-purpose transformer model. This model requires the datasets to be unstylized, in English, and in lowercase. So, the process of cleaning the data in the dataset was implemented using preprocessing where the posts were converted to lowercase. The proposed model trains itself bidirectionally. This means that each word is not just vectorized, but also understood and analyzed by the model in the context of its sentence and the sentiment expressed.

To prevent the model from becoming overly biased toward certain topics, regularization was done and trained in small batches of 32 before reweighting. This enhances the efficacy of our model, particularly in the context of noisy datasets.

Forum for Information Retrieval Evaluation, December 17-20, 2025, Varanasi, India
*Wordson Robert
*Corresponding author.
†These authors contributed equally.
jerinmahibha@msec.edu.in (C.Jerin Mahibha); wordsonrobert@gmail.com (W. Robert); gshimi2022@gmail.com (G. Shimi); theni_d@ssn.edu.in (D. Thenmozhi)

The paper is organized with Section 2 and 3 discussing on related works and datasets, Section 4 on system description, Section 5 discussing results, Section 6 and 7 contributing the error analysis and conclusion.

2. Related Works

The classification of texts as paraphrase or not was determined using the probability associated with the two sentences which is also a classification task. The embedding technique used was GloVe (Global Vectors for Word Representation). GloVe is an unsupervised learning algorithm in which the model trains itself on words and their contexts, drawn from a much larger corpus than the dataset itself. The downsampling technique used here checks the similarity between only the highest-valued tokens and ignores the rest. They had allowed more variation in sentence structure and vocabulary, while focusing on capturing larger, more significant semantic patterns [1].

Klaus et al. [2] had introduced a new method for analyzing legal regulations for individuals without a legislative background. The method had used ROUGE overlap to create concise, representative summaries of the original legal documents. This had made it easier to identify correlations with the relevant laws on which the model had been trained. ROUGE overlap was the core mechanism of this process, guiding the selection of sentences that most closely resembled human-written summaries. This ensured that the resulting summaries are contextually appropriate and accessible to non-experts. Additionally, an imbalanced dataset is balanced through random downsampling of sentences with lower selection probabilities, improving the quality of the training data.

The sentiment analysis forms the base for applications where the public views could be known such as social media posts. This paper had show how a multi label classification of the given text could be implemented by considering the sentiment associated with the text. The models that are applied for monolingual sentiment analysis may not provide good results when it is extended for code mixed data. The paper had illustrated the use of Cross Lingual model on the Tamil - English code mixed data, to classify the sentiment associated with the text instances [3].

Excessive use of social media and the negative facets that guide it, can exacerbate or cause impressions of distress. The nonstop exposure to cautiously curated lives, social comparison, cyberbullying, and the pressure to meet unreal standards can impact an individual's pride, social connections, and overall well-being. The model had identified the levels of depression from social media text using different transformer models like ALBERT and RoBERTa [4].

3. Data set

The dataset used to train and evaluate our model was provided by the organizers of the CryptoQA track as part of FIRE 2025. It consists of user-generated content from three major platforms: YouTube, Reddit, and Twitter (now X). The data includes raw, highly conversational posts centered around cryptocurrency. Three subsets were provided, with 5000 YouTube comments, 4288 Twitter tweets, and 5000 Reddit comments.

To prepare the data for training and evaluation, we randomly split each platform-specific subset into two parts, allocating 80 percent for training and the remaining 20 percent for validation. Thus, the model was trained using 4000 instances from YouTube and Reddit platform and 3430 instances from Twitter platform. The remaining 1000 instances from YouTube and Reddit platform and 858 instances from twitter were reserved for evaluation. This split ensures a balanced evaluation across all three sources while preserving the diversity of conversational styles present in the dataset. The test dataset provided by the organizers had 500 instances each under YouTube, Twitter, and Reddit comments. The data dirstibution of training, evaluation and test dataset are shown in Table 1. The graph shown by Figure 1 gives a visual distribution of all the given datasets.

Table 1
Data Distribution

Language	Training Dataset	Evaluation Dataset	Test Dataset
YouTube	4000	1000	500
Reddit	4000	1000	500
Twitter	3430	858	500

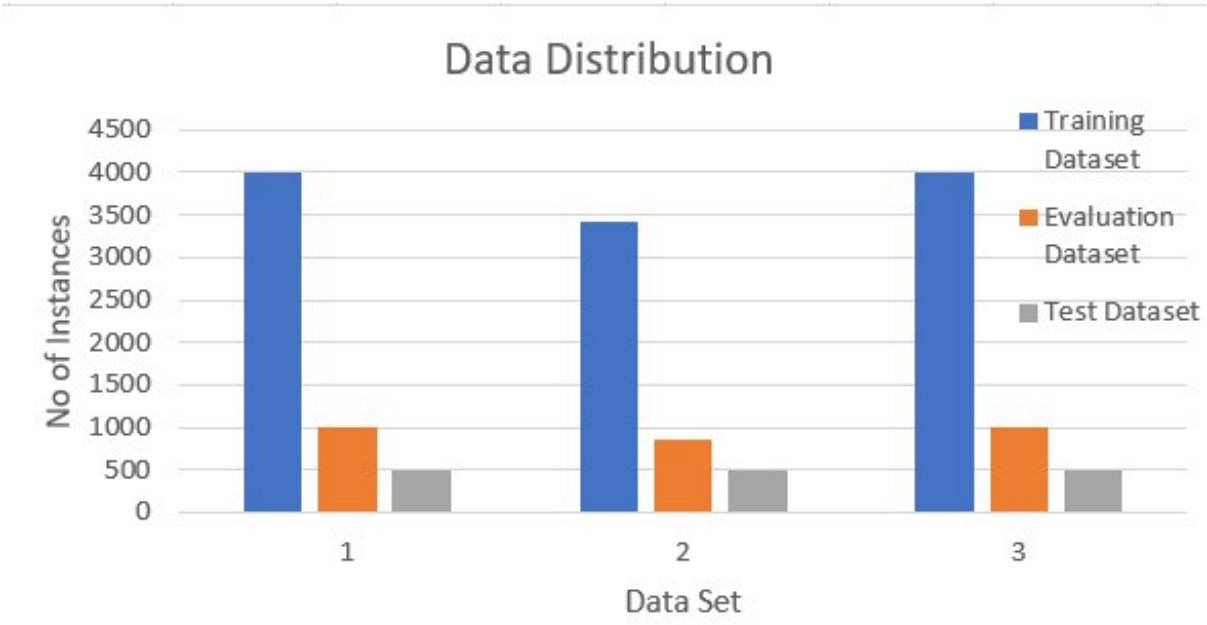

Figure 1: Data distribution

4. System Description

The proposed system's architecture uses an extensively trained, monolingual, lowercase-specific BERT-base model to classify social media posts discussing various cryptocurrency-related topics. Our goal is to contextualize each given sentence using sentiment analysis and classify it into predefined, hierarchically structured categories. The figure below provides a pictorial representation of our model's methodology. The training dataset consists of social media posts collected from platforms such as X (formerly Twitter), YouTube comments, and Reddit. The following subsection discusses the specific details of our model's optimization strategy. After training, we evaluated the model's performance on a separate test set using standard metrics, including accuracy, precision, recall, and F1 score.

4.1. Methodology

Data classification, is the process of categorizing data. In order to accomplish our task of classifying data through sentiment analysis, we had to select an appropriate model from among the many available options. We selected a model that was efficient and well-suited to our specific task to do so effectively. We selected the monolingual, lowercase-specific BERT model because our dataset only contains English text. This model is trained exclusively on lowercased English text and cannot distinguish capital letters. It has 12 transformer layers, each with 12 self-attention heads (for a total of 144 heads), and it uses 768-dimensional token embeddings per position. The model has already been pretrained on large-scale English datasets such as Wikipedia and BookCorpus. To enhance the efficiency of the data classification process, we employed components such as the WordPiece tokenizer, special tokens, and attention masks. Attention masks improve model performance on longer sequences by ensuring computational resources

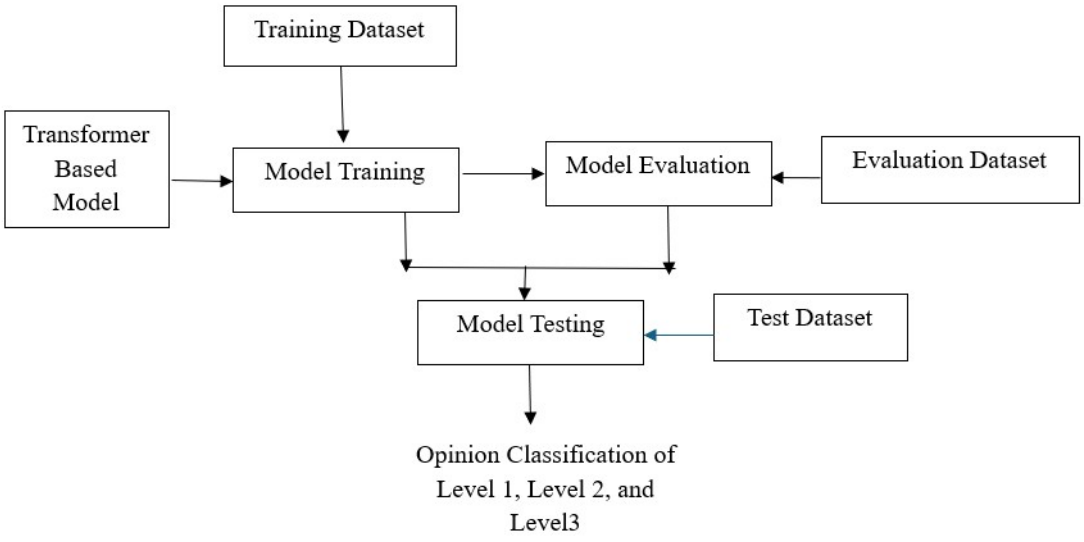

Figure 2: System Architecture

focus on the content rather than the padding. To prevent overfitting, particularly with noisy datasets like ours, we applied dropout regularization, randomly deactivating approximately 30% of the neurons during each training iteration. This prevents the model from becoming too rigid or overly dependent on specific training data. This approach is useful for broad topics, such as cryptocurrency, where many training sentences are only loosely related to the target categories. For more narrowly defined domains, this dropout strategy can be adjusted or omitted. Special tokens, [CLS] and [SEP], were appended to the input sequences to mark the beginning and end of the text, respectively. The model was fine-tuned over four epochs with ten output labels that are structured hierarchically.

Level 1: NOISE, OBJECTIVE, SUBJECTIVE

Level 2 (if SUBJECTIVE): NEUTRAL, NEGATIVE, POSITIVE

Level 3 (if NEUTRAL): NEUTRAL SENTIMENTS, QUESTIONS, ADVERTISEMENTS, AND MISCELLANEOUS.

Thus, the task assigned in the CryptoQA track is a hierarchical classification problem. To address this, a single recursive model architecture was developed instead of using multiple independent flat classifiers. This approach greatly improves the efficiency and accuracy of the overall system. To complement dropout regularization, the AdamW optimizer (Adaptive Moment Estimation with Weight Decay) was used. AdamW improves generalization by smoothing gradients and separating weight decay from gradient updates. These improvements lead to more stable predictions, faster convergence, and better performance during training. These design decisions together contributed to a reduction in cross-entropy loss and helped minimize class imbalance. Increasing the number of layers or embedding dimensions allows the model to capture deeper sentiment cues and recognize a broader spectrum of semantic features, which is necessary for understanding crypto-related social media discourse. 3.

5. Results and Discussion

To assess the accuracy of the proposed model, the parameter macro-F1 score has been used by the organizers. Macro-F1 is a very balanced metric to use because it gives equal importance to how well we predict each of the labels, even if the sample sizes for them are unequal. It's the unweighted average of the F1 scores for each label in the dataset. And the F1 score itself is the harmonic mean of how many correct predictions we made for a class label and how many actual instances of that class labels were

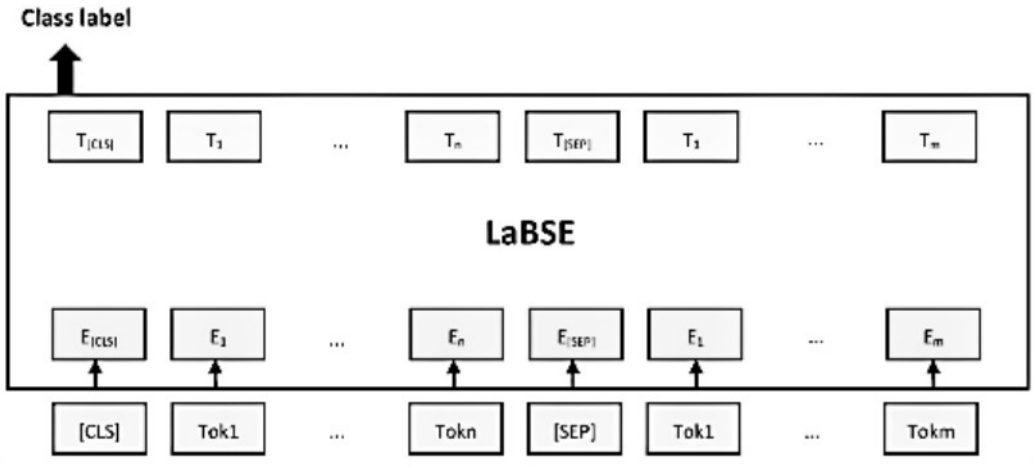

Figure 3: Language Agnostic BERT Model

Table 2
Performance score

Dataset	Level	F1 Score
Reddit	Level 1	0.4709
Reddit	Level 2	0.1053
Reddit	Level 3	Empty
Twitter	Level 1	0.3832
Twitter	Level 2	0.2632
Twitter	Level 3	Empty
YouTube	Level 1	0.1326
YouTube	Level 2	0.1449
YouTube	Level 3	Empty

found correct.

The performance scores of the proposed models have been represented in the table 2. The proposed model turned up with a macro F1 Score of 0.4709 for Level 1, 0.1053 for Level 2 and Empty set for Level 3. Considering Twitter the scores were 0.3832, 0.2632 and Empty set for Level 1, 2 and 3. For YouTube comments the scores were 0.1326 for Level 1, 0.1449 for Level 2 and Empty set for Level 3. The average score of the proposed model was 0.25 which was very low compared to the other findings. These scores show that the proposed models do not provide a good performance for the given dataset.

6. Error Analysis

The model's overall low accuracy suggests a significant number of false positives and false negatives. One possible contributing factor is the distinct styling and language patterns present on different social media platforms. The steep decline in performance from Level 1 to Level 2 is concerning. The complete lack of classifications at Level 3 is also concerning. These issues highlight limitations in the proposed model. However, consistent accuracy was observed at Level 1, suggesting that specific words or phrases are not merely being memorized by the model. Rather, patterns are being identified and differentiated based on the data by the model. This suggests that the issue primarily lies in the generalization of learned representations to more complex, hierarchical levels of classification. Future work should focus on refining the model to better handle multi-level classification tasks and improve generalization across diverse textual inputs. The occasional mismatches are also areas that can still be improved upon.

7. Conclusion

The categorization of digital currency discussions on the web has become more and more essential for businesses, financiers, and government agencies, allowing for the examination of public opinion, the forecasting of market movements, and the creation of more customized strategic plans. A model trained on crypto-specific datasets is valuable because it can capture the nuances of the unique terminology and context of the cryptocurrency domain. In this study, we participated in the CryptOQ shared task, CoDA 2025. This task involved classifying a curated dataset from various online forums into the following categories: negative, positive, objective, subjective, neutral, and miscellaneous. The evaluation results show that Level 1 classifications achieved moderate F1 scores, but performance dropped a lot for Level 2, and Level 3 wasn't addressed. This shows that there's a lack of generalization and severe limits in the current approach. The low overall accuracy and F1 scores underscore the system's unreliability for broader practical applications. These outcomes demonstrate that despite some success in simpler classification tasks, the model struggles to maintain performance across higher-level, more complex categories. Therefore, the focus of future work must be on improving generalization, expanding training data, and refining model architectures to address these challenges and enhance the robustness of crypto-related online conversation classification.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] A. Razaq, Z. Halim, A. Ur Rahman, K. Sikandar, Identification of paraphrased text in research articles through improved embeddings and fine-tuned bert model, Multimedia Tools and Applications 83 (2024) 74205–74232.

[2] S. Klaus, R. Van Hecke, K. Djafari Naini, I. S. Altingovde, J. Bernabé-Moreno, E. Herrera-Viedma, Summarizing legal regulatory documents using transformers, in: Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval, 2022, pp. 2426–2430.

[3] J. M. C, K. Sampath, D. Thenmozhi, Sentiment analysis using cross lingual word embedding model., in: FIRE (Working Notes), 2021, pp. 1094–1100.

[4] M. M, J. M. C, T. D., TechWhiz@LT-EDI-2023: Transformer models to detect levels of depression from social media text, in: B. R. Chakravarthi, B. Bharathi, J. Griffith, K. Bali, P. Buitelaar (Eds.), Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion, INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, 2023, pp. 198–203. URL: https://aclanthology.org/2023.ltedi-1.30/.