Navigating Crypto Conversations: A Multi-Level Approach to Classify Cryptocurrency Opinions on Social Media

Kavya G[1,*], Sonith D[2] and H L Shashirekha[3]

Department of Computer Science, Mangalore University, Mangalore, Karnataka, India

Abstract

In the rapidly evolving cryptocurrency market, understanding public sentiments is pivotal for grasping market trends. People express their views on cryptocurrencies across a variety of social media platforms like Twitter, Reddit, Facebook, and various other online forums. Analyzing sentiments on social media channels makes it necessary to provide a more nuanced perspective on public opinion. However, this task is complicated by the inherent diversity and often unstructured nature of social media content. Posts on social media usually differ greatly in format, tone, and language, which poses significant challenges for accurate sentiment classification. In this direction, "CryptOQA - Understanding CryptoCurrency Related Opinions and Questions from Social Media Posts" - a shared task organized at Forum for Information Retrieval Evaluation (FIRE) 2024, invites the research community to address the challenges of identifying the category of opinions published on social media related to cryptocurrencies in English language. To address these challenges, in this paper, we - team MUCS, describe the models proposed for Task 1: "Opinion Classification from CryptoCurrency related Social Media Posts" of the shared task. We submitted two models: i) Unique_Label_LSTM - a Long Short-Term Memory (LSTM) model with a unique labeling concept and ii) HCC_LSTM - a Hierarchical Classifier Chain (HCC) using LSTMs, to classify the given unlabeled English Reddit and Twitter cryptocurrency opinion texts into one of the predefined hierarchical categories. Among the submitted models, HCC_LSTM obtained macro F1 scores of 0.574 and 0.328 for Twitter and Reddit opinion posts, securing 4$^{\text{th}}$ and 5$^{\text{th}}$ ranks respectively.

Keywords

Cryptocurrency, Opinion Classification, Unique Labeling, Hierarchical Classifier Chain, Machine Leaning and Deep Learning

1. Introduction

Cryptocurrency represents a revolutionary shift in the financial landscape, characterized by digital or virtual currencies that leverage cryptographic techniques for secure transactions. Unlike traditional currencies issued by governments, cryptocurrencies operate on decentralized networks based on blockchain technology. Major cryptocurrencies such as Bitcoin, Ethereum, and Litecoin have gained widespread recognition and adoption, becoming significant players in global financial markets. As these digital assets grow in prominence, so does the importance of understanding public sentiments/opinions surrounding them.

Social media platforms have emerged as vital sources of information and opinion, offering real-time insights into public attitudes and behaviors. Platforms such as Twitter, Reddit, and Facebook, serve as venues where users actively discuss and share their perspectives on various topics, including cryptocurrencies [1]. These discussions can significantly influence market perceptions and trends, making it crucial to monitor and analyze the opinions expressed in these social media conversations. Classifying the sentiments about cryptocurrency related social media posts provides valuable insights into public opinion and can help stakeholders to make informed decisions. Accurate classification of opinions ranging from positive and negative to neutral and objective can assist investors in predicting market movements, understanding consumer behavior, and shaping marketing strategies [2]. However, the task of opinion classification in this context is fraught with challenges due to the diverse and often unstructured nature of social media content. One of the primary challenges is the inherent variability

Forum for Information Retrieval Evaluation (FIRE), December 12-15, 2024, India

*Corresponding author.
kavyamujk@gmail.com (K. G); sonithksd@gmail.com (S. D); hlsrekha@mangaloreuniversity.ac.in (H. L. Shashirekha)
https://www.mangaloreuniversity.ac.in/shashirekha (H. L. Shashirekha)

and ambiguity of social media language. Posts on social media are usually short, informal, and laden with slang, making it difficult for traditional text classification models to accurately interpret the hidden sentiments in the given text [3]. Additionally, the context in which opinions are expressed can vary widely, further complicating the classification process. These issues necessitate the use of sophisticated models that can handle the complexities of social media language and classify the opinions more accurately.

To address the challenges of identifying cryptocurrency related opinion posts in social media platforms, "CryptOQA Understanding CryptoCurrency related Opinions and Questions from Social Media Posts" shared task[1] organized at FIRE[2] 2024, invites the research community to develop models to detect the categories of cryptocurrency posts. The shared task consists of two tasks: Task 1 - Opinion Classification from CryptoCurrency related Social Media Posts and Task 2 - Question Answering from CryptoCurrency related Social Media Posts. We participated in only Task 1 which is about the classification of Twitter and Reddit cryptocurrency related social media opinion posts. The dataset provided by organizers of the shared task consists of 5,000 posts per platform, annotated at three levels of hierarchy as shown in Figure 1.

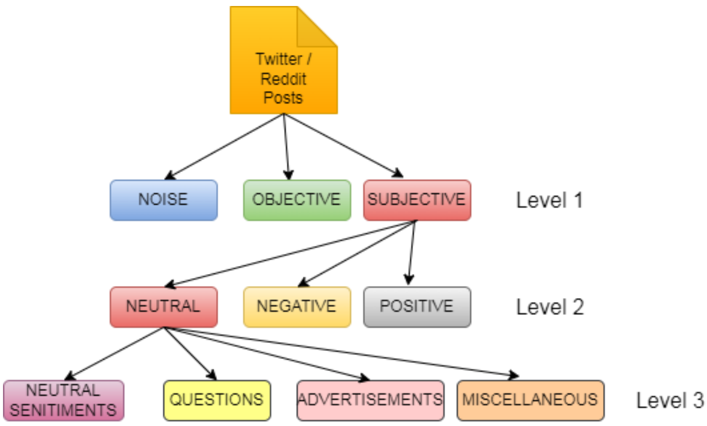

Figure 1: Hierarchy of classes for Task 1: Opinion Classification from CryptoCurrency related Social Media Post

Many real-world scenarios require hierarchical classification, despite the fact that most of the research is focused on flat classification problems. As the classes are arranged in a hierarchy in hierarchical classification, the models have to learn the dependencies between classes in the hierarchy to predict the class label for the unseen instance. Hierarchical classification becomes important when data is organized in multiple levels satisfying the class and sub-class relationship. The challenges of hierarchical classification is usually addressed through strategies such as Breadth First Search (BFS) and Depth First Search (DFS), incorporating contextual information from higher to lower levels in the hierarchy. To explore the strategies of detecting cryptocurrency opinion posts in English on social media platforms, in this paper, we - team MUCS, describe the models submitted to Task 1 of the shared task. We implemented two learning models: i) Unique_Label_LSTM - a LSTM model with a unique labeling concept and ii) HCC_LSTM - a HCC using LSTMs, to tackle the nuances of opinion classification. By leveraging these models, we aim to enhance the accuracy of sentiment analysis and provide more robust insights into public opinion regarding cryptocurrencies.

The rest of paper is organized as follows: Section 2 describes the recent literature on cryptocurrency related opinion mining and Section 3 focuses on the description of the proposed models followed by the experiments and results in Section 4. The paper concludes with future works in Section 5.

[1]https://sites.google.com/view/cryptoqa-2024
[2]https://fire.irsi.org.in/fire/2024/home

2. Related Work

Cryptocurrency market has evolved in recent years, presenting unique challenges and opportunities for analysis. Understanding public sentiments and predicting market trends are critical for navigating this volatile landscape. A range of studies have explored different learning approaches to address these challenges, particularly in the context of sentiment analysis in social media posts related to cryptocurrency. Some notable works are described below:

Aslam et al. [4] explored a combination of Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, utilizing Term Frequency-Inverse Document Frequency (TF-IDF) of words, Bag of Words (BoW) and Word2Vec, for the detection of sentiments and emotions in cryptocurrency related tweets. LSTM-GRU model trained with BoW features achieved 99% and 92% accuracies for sentiment analysis and emotion prediction, respectively. Kim et al. [5] proposed a novel framework of Self-Attention-based Multiple LSTM (SAM-LSTM) model, for predicting Bitcoin (BTC) prices by leveraging a change point detection technique to segment time-series data for improved normalization. When tested on real-world BTC price data, their model achieved notable results with a Mean Absolute Error (MAE) of 0.3462, Root Mean Square Error (RMSE) of 0.5035, Mean Squared Error (MSE) of 0.2536, and Mean Absolute Percentage Error (MAPE) of 1.3251, demonstrating its effectiveness in BTC price prediction. To study the influence of social media like Twitter on cryptocurrencies, Sahal [6] proposed integrating Bidirectional LSTM (BiLSTM) networks with Embeddings from Language Models (ELMo), to identify the most profitable cryptocurrencies by contextually-based sentiment analysis. Their proposed approach achieved an accuracy of 86.30%, demonstrating its effectiveness in predicting price changes based on social media data. Huang et al. [7] focuses on predicting cryptocurrency price fluctuations by analyzing sentiments from social media, particularly from Sina-Weibo, a major Chinese platform. The authors proposed a novel approach that includes capturing Weibo posts, creating a crypto-specific sentiment dictionary, and utilizing LSTM based Recurrent Neural Network (RNN) and auto regression models in conjunction with historical price data. Their proposed LSTM model outperformed auto-regressive models with precision and recall of 0.87 and 0.94 respectively.

To investigate whether Twitter data related to cryptocurrencies can enhance trading strategies for Bitcoin, Colianni et al. [8] proposed Machine Learning (ML) techniques (Bernoli Naive Bayes (BNB), Logistic Regression (LR), Multinominal NB (MNB), and Support Vector Machines (SVM)) trained with binary BoW features. Among their proposed models, BNB classifier achieved the highest accuracy with a day-to-day prediction accuracy of 95% and an hour-to-hour accuracy of 76.23% respectively. Torba et al. [9] presented Hierarchical Text Classification (HTC) as a generative task using advanced neural models consisting of open framework that facilitates experimentation with various aspects of HTC models including traversal strategies for class trees (BFS vs. DFS, root-to-leaf vs. leaf-to-root), constraints on hierarchy coherence during decoding, and the use of label names versus acronyms. This work provides datasets, metrics, and tools for error analysis, enabling researchers to test these modeling choices comprehensively. By evaluating these factors, the authors aims to clarify how different architectural and modeling decisions influence HTC outcomes and promote transparency and reproducibility in the field of classification. Hua et al. [10] proposed Hierarchical Graph Neural Network models (Text Graph Convolutional Network (TextGCN), Simple Graph Convolution (SGC), Text Level GCN (TextLevelGCN), and Hierarchical Graph Attention neural network (HieGAT)) to enhance text classification by effectively leveraging word-level, sentence-level and document-level features. Experimenting on various benchmark datasets (20NG, R8, R52, Ohsumed, and MR), their proposed HieGAT model outperformed other models.

The related work illustrates that researchers have employed various approaches such as LSTM-GRU ensembles, self-attention mechanisms, and hierarchical graph neural networks using diverse features and data sources, to analyze cryptocurrency sentiments. While these models have demonstrated their effectiveness, the evolving nature of cryptocurrency markets and social media content suggests there is still considerable scope for further research and innovation.

3. Methodology

Hierarchical classification organizes labels into a multi-level structure, enabling more detailed and context-aware categorization by considering the relationships between different levels of labels. The proposed methodology includes pre-processing the given data followed by model building as described in the following sub-sections:

3.1. Pre-processing

Pre-processing is an essential step that converts raw text data into a clean and organized format, thereby improving the effectiveness and accuracy of ML models. This process is necessary because raw text data often contains noise, inconsistencies, and irrelevant information that can hinder model accuracy and efficiency. Pre-processing standardizes the data, making it consistent and easier for the models to understand and analyze. In this study, numeric information is transformed into words and URLs, user mentions, hashtags, special characters, punctuation are removed and NaN values are filled with empty strings. Additionally, stopwords are removed using resources available at the Natural Language Toolkit[3] (NLTK) to ensure the text is effectively prepared for analysis. For Reddit dataset, 'title' and 'selftext' fields are concatenated before applying pre-processing.

3.2. Model Building

LSTM networks are a type of RNN which excel at capturing long-term dependencies and understanding contextual relationships within text data [11]. It is a powerful model designed to handle sequential data, making it particularly effective for text processing tasks such as sentiment analysis. LSTM architecture includes Input layer, Embedding layer, LSTM layer, Dense layer that serves as output layer, and an optional Dropout layer. While the Input layer accepts text in sequence format, Embedding layer transforms the text into a dense vector representation using Keras embeddings[4] [5]. This involves using Keras tokenizer to convert text data into sequence of integers, where each unique word is mapped to a distinct integer facilitating numerical representation for the given text. These sequence of integers are then padded to a uniform length, ensuring that all input sequences are of the same length for consistent model input. Further, the integer sequences are mapped to continuous vectors that encode semantic relationships between words which results in dense, fixed-size vector representations. This is followed by 64 units LSTM layer, which processes the embedded sequences to capture dependencies over long text sequences. Further, LSTM network includes two Dense layers: one with 32 units and ReLU activation for introducing non-linearity, and another with a softmax activation function to output class probabilities for the final classification. Dropout layer helps to prevent overfitting by randomly setting a fraction of the input units to zero during training. The model is compiled with the Adam optimizer and sparse categorical cross-entropy loss, trained over 5 epochs with a batch size of 32.

The two proposed models: i) Unique_Label_LSTM and ii) HCC_LSTM, make use of LSTM model to learn relationship between the classes arranged in a hierarchy. While Unique_Label_LSTM make use of only one LSTM as flat classifier, HCC_LSTM use three LSTMs as chain of classifiers. Further, the models differ in the way labels are considered to build the models. The description of building the models is given below:

- **Unique_Label_LSTM model** - unique labels are generated by concatenating the labels in the path from the root of the subtree to each leaf node from level 1 to level 3 in the hierarchy of labels shown in Figure 1. This concept results in eight unique labels - NOISE, OBJECTIVE, SUBJECTIVE_NEUTRAL_SENTIMENTS, SUBJECTIVE_NEUTRAL_QUESTIONS, SUBJECTIVE_NEUTRAL_ADVERTISEMENTS, SUBJECTIVE_NEUTRAL_MISCELLANEOUS, SUBJECTIVE_NEGATIVE and SUBJECTIVE_POSITIVE, for the given hierarchy of labels. The labels

[3]https://www.nltk.org/
[4]https://www.tensorflow.org/api_docs/python/tf/keras/layers/Embedding
[5]https://keras.io/2.15/api/layers/core_layers/embedding/

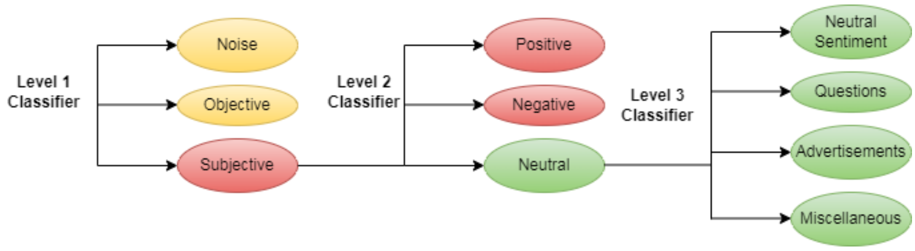

Figure 2: Classifier chains in HCC_LSTM using LSTM model

’NOISE’ and ’OBJECTIVE’ are used directly as they appear in level 1 and do not have branches further. This labeling arrangement helps to build a flat multi-class classifier with eight labels which allows for nuanced classification of opinions, ensuring the systematic assignment of each unseen input to its most relevant category based on its content and context.

- **HCC_LSTM model** - employs a hierarchical approach with multiple local classifiers dedicated for each level of the hierarchy, as shown in Figure 2. For each level, the model is trained using data filtered by the predictions of the previous level, allowing it to refine its classifications hierarchically resulting in a robust framework for hierarchical text classification.

While Unique_Label_LSTM model is a flat classifier, HCC_LSTM is a hierarchical classifier and these classifiers can achieve accurate and context-aware predictions for the new unseen text.

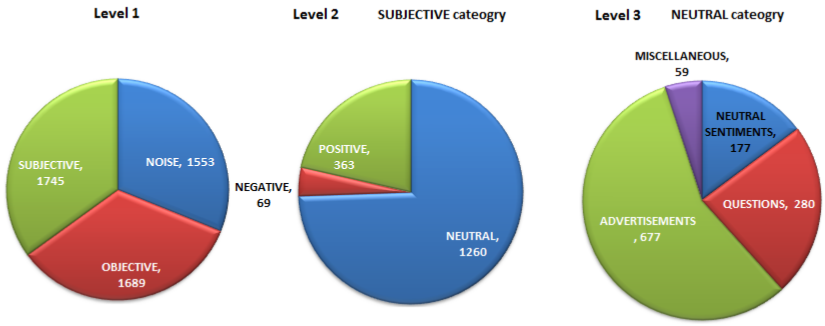

Figure 3: Level-wise distribution of data in Twitter cryptocurrency opinion classification dataset

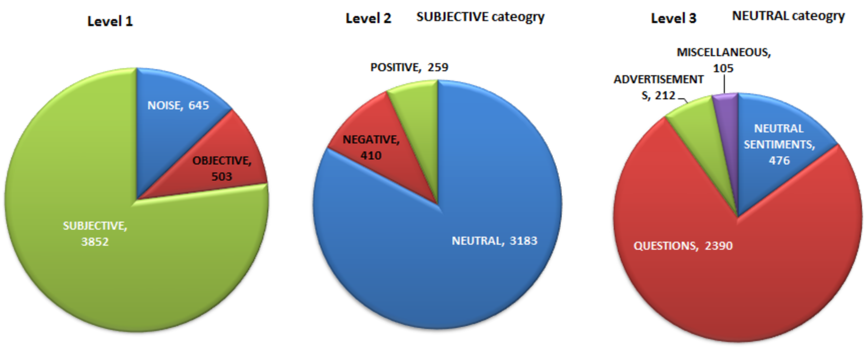

Figure 4: Level-wise distribution of data in Reddit cryptocurrency opinion classification dataset

4. Experiments and Results

Reddit and Twitter cryptocurrency opinion classification datasets provided by the organizers of the shared task consists of only Train sets and the Train sets are imbalanced. The level-wise distribution

Table 1
Sample text and corresponding labels in Twitter cryptocurrency opinion classification dataset

Sample Text	Category		
	Level 1	Level 2	Level 3
start with what is right rather than what is acceptable. -franz kafka #motivation #inspiration #dogecoin #doge #doonlygoodeveryday	0		
rt @ixfiworld: women empowerment in crypto discover how the crypto ecosystem is transforming women is leadership in business	1		
rt @like_minds01: new article update want to earn high returns through yield farming? we have compiled a list of top yield farming protoâ€¦	2	0	0
rt @terra_army: will #bitcoin move upwards from here? if yes then luna can become a rocket. like and rt support $lunc #â€¦	2	0	1
rt @youngsternft: it looks like you liked the last giveaway so now i'm giving away $100 (in eth) to three people to enter: - follow meâ€¦	2	0	2
rt @fantohmdao: happy sunday #fhmily! are you still at your computer? take this time to go outside and experience life outside of #cryptoâ€¦	2	0	3
rt @soldnever: @mattwallace888 i exchanged all of my #bitcoin for #dogecoin april 9th-13th last year but elon musk will be the 1st person tâ€¦	2	1	
this is a very cool and amazing project all the partners of this project are so hardworking and this team has great potential to succeed in this project. #crypto #cryptocurrecy #cryptotwitter #cryptoworld #cryptomemes #memes2022 #cryptomoney #memesdaily #catmemes #catsoftwitter	2	2	

of data in the given Twitter and Reddit datasets are shown in the Figure 3 and 4 respectively and the sample text in Twitter and Reddit datasets are shown in the Tables 1 and 2 respectively. As the datasets consists of only Train sets, 20% of the Train sets at random are considered as Validation sets to evaluate the performances of the models and the remaining as Train sets. The performances of the proposed models evaluated on the Validation set based on macro F1 score are shown in Table 3 for both Twitter and Reddit datasets.

As the shared task participants were allowed to submit the predictions of only two models on the Test sets, we trained the models with the given Train sets and obtained the predictions on the Test sets provided by the organizers. These predictions are evaluated by the organizers based on macro F1 score and the proposed HCC_LSTM model obtained better macro F1 scores of 0.574 and 0.328 for Twitter and Reddit datasets, securing 4^{th} and 5^{th} ranks respectively compared to the other proposed model. Comparison of macro F1 scores of all participating teams are shown in Figures 5 and 6 respectively.

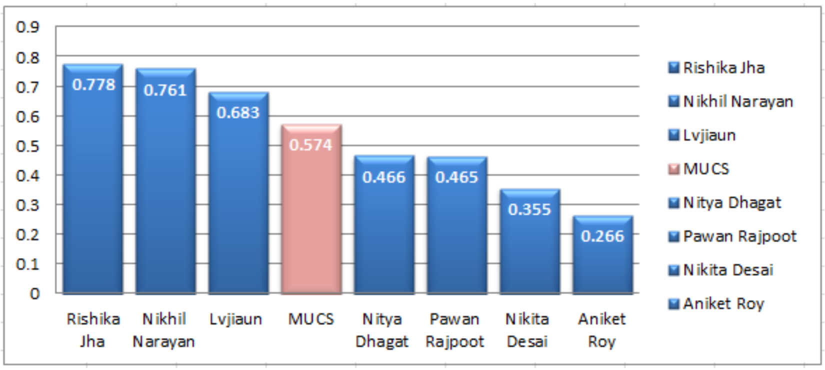

Figure 5: Comparison of macro F1 scores of all participating teams in Task 1: Twitter opinion classification

Table 2
Sample text and corresponding labels in Reddit cryptocurrency opinion classification dataset

Title	Selftext	Category		
		Level 1	Level 2	Level 3
Cooler Master PCIE Cables SWAP	i have cooler master v850 gold v2 and cooler master mwe 1050w gold v2. i burnt the 1050w pcie cables and i'm thinking of using the 850w pcie cables on the 1050w psu. Any advice?	0		
Russia to adopt crypto as legal tender	Works are under way in Russia to recognise BTC and Alt Coins as legal tender. More details can be found [here] (https://threadreaderapp.com/thread/1491313537092034563.html)	1		
Anyone wants to Sell Bitcoin on Vauld and get money via UPI.	Vauld isn't accepting UPI or Bank payments but I wanted to add a few thousands worth Bitcoin.\n\nDM me if anyone wants to sell their Bitcoin, we can transfer Bitcoiy through Vauld transfer system and I can pay via UPI to their bank Accounts.	2	0	0
How does TDS work ?	So, I was selling USDT on paxful and saw a lot of people in their offer details written ?PAN required?, ? 1% TDS will be deducted ? etc .\n\nhow does it work ? they give our 1% TAX with our PAN and it will show on our bank that we traded crypto? idk how that works can someone explain ?	2	0	1
New-Crypto-Themed Discord Channels	Hi, I've been dedicating the whole week to finding new cryptocurrencies or tokens and buying the ones that look very decent. For example I bought SafeBTC and it grew x4 it's original price. \n\nI was wondering if you guys do the same and talk about it on your discord channels, and if you do, could you please leave a link below so we can all join? \n\nThank u so much in advance, and love the passion of the community.	2	0	2
Elon Musk: The Fiat Rockefeller	If you guys haven't had a chance to listen to this episode, I highly recommend it. It's a great take on Musk's lack of understanding of what Bitcoin truly is. \n\n[https://saifedean.com/podcast/57-elon-musk-the-fiat-rockefeller/] (https://saifedean.com/podcast/57-elon-musk-the-fiat-rockefeller/)	2	0	3
Warning to KuCoin Users	Do yourself a favor and don't use Kucoin\n\nThey are not reliable, I transferred my coins from Coinbase Pro to KuCoin and my coins are now lost\n\nTheir support doesn't respond to my emails.\n\nI have verified the address is correct and i have verified that the transaction was successful through https://etherscan.io/	2	1	
FUD will turn to FOMO	To those that are new to crypto....donï¿½t be the person that is to scared to buy when bitcoin is above 50k saying itï¿½s too high then doesnï¿½t buy now when bitcoin is on sale. You will be the same person buying at 80k\n\nBe wise and just keep buying	2	2	

Table 3
Performances of proposed models on the Validation set of Twitter and Reddit datasets

Model	**Twitter**				**Reddit**			
	P	**R**	**MF1 score**	**WF1 score**	**P**	**R**	**MF1 score**	**WF1 score**
Unique_Label_LSTM	0.494	0.520	0.506	0.717	0.347	0.319	0.323	0.562
HCC_LSTM	0.805	0.694	0.728	0.805	0.729	0.686	0.703	0.794
P: Precision; R: Recall; MF1: Macro F1 score; WF1: Weighted F1 score								

5. Conclusion and Future Work

In this paper, we - team MUCS, describe the models submitted to Task-1: 'Opinion Classification from CryptoCurrency related Social Media Posts' of the shared task "CryptOQA Understanding CryptoCurrency related Opinions and Questions from Social Media Posts" at 'FIRE 2024', to distinguishing between categories of cryptocurrency related Twitter and Reddit opinion posts in English. We submitted two models: i) Unique_Label_LSTM - a LSTM model with a unique labeling concept and ii) HCC_LSTM - a HCC using LSTMs, to classify the given unlabeled English Reddit and Twitter opinion texts into one of the predefined hierarchical categories. Among the submitted models, HCC_LSTM obtained macro F1 scores of 0.574 and 0.328 for Twitter and Reddit opinion posts, securing 4^{th} and 5^{th} ranks respectively.

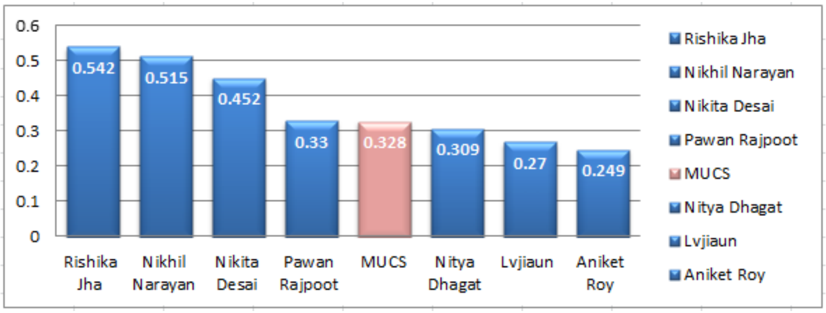

Figure 6: Comparison of macro F1 scores of all participating teams in Task 1: Reddit opinion classification

Investigating other approaches to capture more nuanced opinions about cryptocurrencies from social media data to improve the performance of the learning models will be explored further.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to: Grammar and spelling check. 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] K. R. K. G. Sougata Sarkar, Gourav Sen, 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.

[2] M. Khan, et al., Predicting Cryptocurrency Value, Based on Sentimental Analysis of Social Media Post, 2022.

[3] M. de Guerra Narciso, Cryptocurrency Analysis Based on User-Generated Social Media Content, 2020.

[4] 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, in: Ieee Access, volume 10, IEEE, 2022, pp. 39313–39324.

[5] G. Kim, D.-H. Shin, J. G. Choi, S. Lim, A Deep Learning-based Cryptocurrency Price Prediction Model that Uses On-chain Data, in: IEEE Access, volume 10, IEEE, 2022, pp. 56232–56248.

[6] R. Sahal, Predicting Optimal Cryptocurrency using Social Media Sentimental Analysis, 2023.

[7] X. Huang, W. Zhang, X. Tang, M. Zhang, J. Surbiryala, V. Iosifidis, Z. Liu, J. Zhang, Lstm Based Sentiment Analysis for Cryptocurrency Prediction, in: Database Systems for Advanced Applications: 26th International Conference, DASFAA 2021, Taipei, Taiwan, April 11–14, 2021, Proceedings, Part III 26, Springer, 2021, pp. 617–621.

[8] S. Colianni, S. Rosales, M. Signorotti, Algorithmic Trading of Cryptocurrency Based on Twitter Sentiment Analysis, in: CS229 Project, volume 1, 2015, pp. 1–4.

[9] F. Torba, C. Gravier, C. Laclau, A. Kammoun, J. Subercaze, A Study on Hierarchical Text Classification as a Seq2seq Task, in: European Conference on Information Retrieval, Springer, 2024, pp. 287–296.

[10] S. Hua, X. Li, Y. Jing, Q. Liu, A Semantic Hierarchical Graph Neural Network for Text Classification, in: arXiv preprint arXiv:2209.07031, 2022.

[11] A. Rao, N. Spasojevic, Actionable and Political Text Classification Using Word Embeddings and LSTM, in: arXiv preprint arXiv:1607.02501, 2016.