CryptOpiQA: A new Opinion and Question Answering dataset on Cryptocurrency

Sougata Sarkar[1], Aditya Badwal[2], Amartya Roy[3], Koustav Rudra[4] and Kripabandhu Ghosh[5]

[1]Deloitte *, sougata.sarkar8101@gmail.com

[2]Indian Institute of Science Education and Research Kolkata, adityabadwal@gmail.com

[3]Bosch, Amartya.Roy@in.bosch.com

[4]Indian Institute of Technology Kharagpur, krudra@ai.iitkgp.ac.in

[5]Indian Institute of Science Education and Research Kolkata, kripaghosh@iiserkol.ac.in

Abstract

Cryptocurrency has attracted a lot of public attention and opinion worldwide. Users have different kinds of information needs regarding such topics and publicly available information is a good resource to satisfy those information needs. In this paper, we investigate the public opinion on cryptocurrency and bitcoin on two social media – Twitter and Reddit. We have created a multi-level dataset *CryptOpiQA* and garnered valuable insights. The dataset contains both gold standard (manually annotated) and silver standard (inferred from the gold standard) labels. As a part of this dataset, we have also created a Question Answering sub-corpus. We have used state-of-the-art LLMs and advanced techniques such as retrieval augmented generation (RAG) to improve question-answering (QnA) results. We believe this dataset and the analysis will be useful in studying user opinions and Question-Answering on cryptocurrency in the research community.

1 Introduction

Cryptocurrency has gained a lot of public interest in recent years. Given the potential of cryptocurrency to be a digital substitute for fiat currency (e.g. U.S. dollar, the British pound, the Indian rupee, and the Euro), different governments have different stances on legalizing cryptocurrency (Inv, 2023).

While works have focussed on creating blockchain networks (Nguyen et al., 2022; Messias et al., 2023) or leveraging social media for tasks like price/market prediction, locating bounty events (Tandon et al., 2021; Abraham et al., 2018; Karolis Zilius, 2023; Ali Raheman, 2022), to our knowledge, there exists no study on the fine-grained public opinion on cryptocurrency. The shared task (CLE, 2023) is focussed on user profiling on cryptocurrency on Tweets on a few training sets. In addition, question-answering on cryptocurrency over social media is also an interesting task given the ubiquitous inquisitiveness on cryptocurrency.

In this paper, we create a dataset (named **CryptOpiQA**)[1] consisting of i) Tweets from the U.S. ii) Reddit posts concerning cryptocurrencies on multi-level fine-grained classes (the Opinion dataset), iii) question-answering dataset derived from the main Tweet dataset and iv) question-answering dataset derived from the main Reddit dataset (the QnA dataset). For both social media (Twitter and Reddit), we annotate manually the opinions (on cryptocurrency) exhibited in these posts to generate the *gold standard data* pertaining to a multi-level fine-grained hierarchy and leverage machine learning to generate a high-quality *silver standard data* (see **Section 3** for details). From the gold and silver standard posts (as questions) of the two media, we further create annotated question-answering datasets where the answers of these posts are further curated and annotated (see **Section 3.4**). Furthermore, we ran classification experiments (**Section 4.1**) on the main (opinion) dataset and both unsupervised and supervised ranking experiments on the question-answering dataset and finally used RAG on the QnA dataset(see **Section 4.2**). Finally, we derived insightful analysis on the dataset (See **Section 4.3**).

We observe that identifying opinions in Tweets and Reddit posts is intricate, and it becomes more challenging as we go into more granular levels. In **Section 4.3** we analyze the popular posts on both Twitter and Reddit. We believe our dataset can be leveraged on fine-grained opinion classification on social media. The question-answering dataset, apart from serving as a benchmark, can be a repository of the most interesting questions and the relevant answers thereof, on cryptocurrency. The purpose of this paper is to study public opinion on cryptocurrency – possibly a relatively lesser un-

*The project was done during MS thesis at Indian Institute of Science Education and Research Kolkata

[1]Currently it is available in this zenodo repository: https://doi.org/10.5281/zenodo.14469000

Proceedings of the 31st International Conference on Computational Linguistics, pages 11107–11120
January 19–24, 2025.

derstood topic worldwide and hence with a chance for diverse opinions (educated or branching from unawareness). We thought that mining such opinions from two social media platforms (Twitter and Reddit) could shed some light on the same esp. because, to our knowledge, there exists no publicly available dataset dedicated to this study. In addition, we extract a QnA dataset specifically curated to study the inquisitiveness and within-community responses on these social media platforms. This can reveal interesting user views on cryptocurrencies that can be possibly harnessed by policymakers, economists, and marketing experts.

Phrase	Subreddit	Post	Comments
'btc', 'bitcoin'	bitcoin	274435	3915445
	bitcoinbeginners	50396	316376
	bitcoinindia	4875	7875
	btc	86235	655658
'crypto'	cryptocurrencies	97517	109237
	cryptocurrency	808796	18445864
	cryptoindia	3938	17334
	cryptomarkets	138012	285063
	cryptomoonshots	417024	913084
	cryptotechnology	12118	43183
'ethereum'	ethereum	90332	559274
'ripple'	ripple	16375	104049
Total		2000053	2537244

Table 1: **Data statistics of posts and comments collected from Reddit**

2 Related Work

Cryptocurrency has gained a lot of popularity, mostly due to the increase in return from cryptocurrency and its acceptance as an asset (Cheong and Lin, 2024) in most countries. There is a lot of debate about accepting cryptocurrency as a currency for a country; though some countries have accepted it, some have an implicit ban, and others completely banned crypto as a currency (Alvarez et al., 2022). So, now the question stands as to whether other countries should also accept it. The answer is much more complex than that because every country has its own currency system and taxation on that, but crypto can completely shake that system, and the Government may lose a lot of assets (Marian, 2013), and there is a lot about the price control of the cryptocurrency(Sanshao Peng and Sarker, 2023).

Our study goes in a different direction than the above-mentioned prior work. We instead explore the public opinion encircling cryptocurrency from social media posts. To our knowledge, this has *not* been done before. For the past few decades, social media has emerged as the primary medium for public opinion. We have considered two of the most famous social media: Reddit and Twitter. Text datasets from social media help face many open world problems (Guo et al., 2019; https://github.com/cvjena/twitter-flood-dataset; Poddar et al., 2022). Our dataset tries to portray people's emotions about cryptocurrency obtained from social media. There are already plenty of models and datasets that can detect emotions from texts (Muhammad et al., 2022; Antypas and Camacho-Collados, 2023; Jiménez Zafra et al., 2015) and also detect abusive behavior of a user (Verma et al., 2020; Founta et al., 2018). The main contribution of our dataset stands for novelty, the establishment of the hierarchy (See **Figure 2**) from manual observation of texts, and most importantly, thorough manual annotation maintaining the hierarchy. To our knowledge, this is the last cryptocurrency dataset scraped from Twitter and Reddit using the official API due to restrictions imposed by these media. In the later sections (See **Section 4**), we have showcased the usability of the dataset.

3 Dataset

This paper aims to understand different types of opinions posted on social media about cryptocurrency. First, we explain the data collection procedure followed by a detailed annotation process.

3.1 Data Collection

We collected cryptocurrency-related posts from Twitter and Reddit in May 2022 or later. The collection procedure is as follows:

Twitter: We have used the standard twitter API, tweepy(https://docs.tweepy.org/en/stable/) to collect the tweets. A set of thirty-two hashtags (e.g., #altcoin, #bitcoin, #cryptocurrency, #crypto, #dogecoin, #ethereum etc.) related to cryptocurrency is used to collect the tweets. Additionally, we have used location tags "US" in the Twitter field *country_code* to collect the data from the USA (All the tweets with location tag as the USA assuming them to be posted from USA). Finally, we have collected **33,541** tweets. The reason for choosing the USA is that the USA is one of the few countries that accepts Crypto as a legal currency.

Reddit: We employed the Pushshift API Wrapper (PSAW) (https://psaw.readthedocs.io/en/latest/)

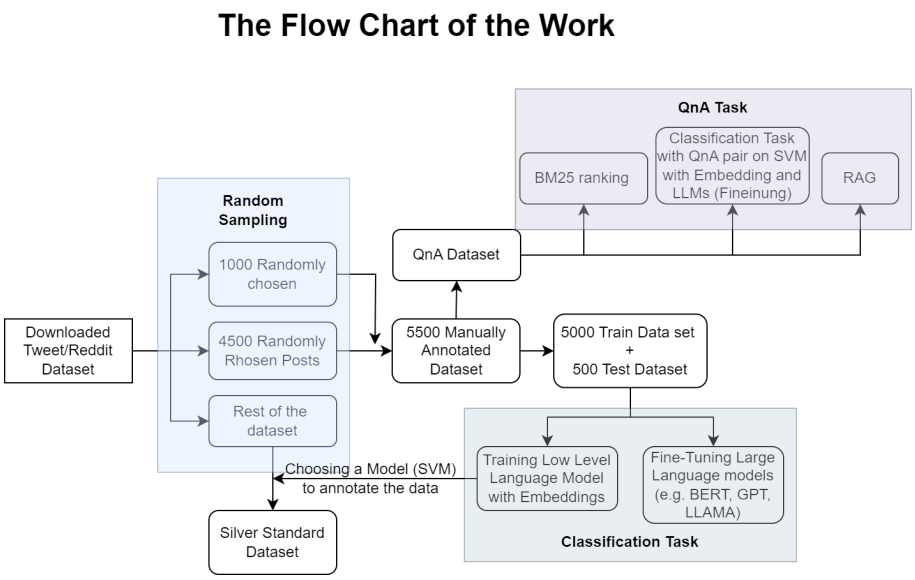

Figure 1: This is the entire workflow of creation of and experiments with the dataset

to gather posts from relevant subreddits on Reddit. PSAW is a Python library that provides an interface to the Pushshift API, a popular tool for accessing historical Reddit data. Subreddits were selected based on their inclusion of terms indicative of cryptocurrency interest, such as “btc”, “bitcoin”, “crypto”, “ethereum”, and “ripple”. Due to the large number of subreddits containing the term "crypto", we focused on those with a high subscriber base: at least 300k members for global subreddits and 100k members for India-specific ones. After identifying ten such subreddits, we retrieved all posts containing the aforementioned keywords using PSAW. Following post-retrieval, we collected comments associated with each post. This was achieved by utilizing a dedicated script[2] designed to interact with the official Reddit API (PRAW). PRAW allows programmatic access to Reddit’s functionalities, enabling the extraction of comments for each retrieved post. **Table 1** summarizes the collected Reddit data statistics.

Through manual analysis, we observed different types of posts on the two aforementioned social media on cryptocurrency. To represent the different fine-grained contents of the posts, we construct the three-level hierarchy as follows:

1. **Level 1:** At this level, the data is classified into three classes, i.e., (i) **Objective**: data containing factual information, (ii) **Subjective**: data carrying opinions/sentiments of users, and (iii) **Noise**: irrelevant content, mostly out of topic, or not well understood w.r.t the texts.

2. **Level 2:** At this level, Subjective tweets are further categorized into three different labels: (i) **Positive**: tweets/Reddit posts that contain positive feedback about cryptocurrencies, (ii) **Negative**: posts expressing negative sentiment about cryptocurrencies, (iii) **Neutral**: these kinds of posts represent the general opinion, questions, etc.

3. **Level 3:** At this level, Neutral subjective posts are further categorized into the following four classes: (i) **Neutral sentiments**: posts carry sentiments that are neither positive nor negative, (ii) **Questions**: users ask questions about cryptocurrency, (iii) **Advertisements**: posts that promote the cryptocurrency investment, (iv) **Miscellaneous**: rest of the posts that can not be classified into above mentioned classes. Here we have put the tweets that are mixed of the above-mentioned classes and also the ones that seemed like scams. (see **Table 2**)

Figure 2 pictorially shows the hierarchy.

3.2 Creation of Gold Standard dataset

In this subsection, we describe the creation of the Gold Standard (manually annotated) dataset on tweets and reddit posts.

Manual annotation of random initial sample: We first randomly sampled 1,000 posts from each platform (Twitter and Reddit) to establish a hierar-

[2] https://github.com/pistocop/subreddit-comments-dl

Category	Source	Example
Noise	tweet	happy independence day, americans.#dogecoin
	reddit	State of this sub right now
Objective	tweet	the current value of 1 doge in usd is: $0.067165 (up 0.001050 so far today). #dogecoin
	reddit	Bitcoin shoots past $ 51,000 adding over $ 70 billion
Negative	tweet	#dogecoin is just barely getting started
	reddit	Trying to manipulate the market once again
Positive	tweet	doge, the only crypto cool enough to board through space #dogecoin
	reddit	Ethereum making big moves in 2020.
Neutral	tweet	there is a lot of misinformation going around about #dogecoin from mockers and doubters
	reddit	About the new tds update.
Questions	tweet	what is polygon and why will non-fungible history be deployed there #cryptocurrency
	reddit	what is the next upcoming crypto coin youre looking into ?
Advertisement	tweet	10.000$ busd giveaway we are giving away 10.000 busd to 20 lucky winners
	reddit	#FREE Automated Crypto Trading Bot , Crypto Signals on Mobile App – Crypto Classifieds
Miscellaneous	tweet	drop your #dogecoin wallet address,100,000 $doge will be sent to your wallet
	reddit	How to Buy Bitcoin: A Step by Step Guide to buy it Fast, Easy Safe

Table 2: Examples of different Tweet/Reddit post classes as shown in the hierarchy in **Figure 2**.

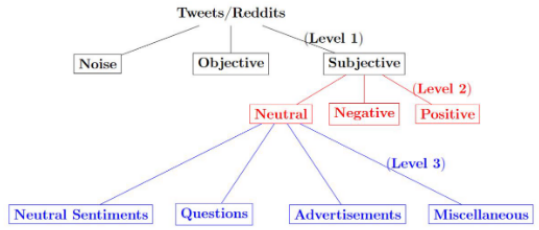

Figure 2: Hierarchy of the Tweet/Reddit classes used for an- notation

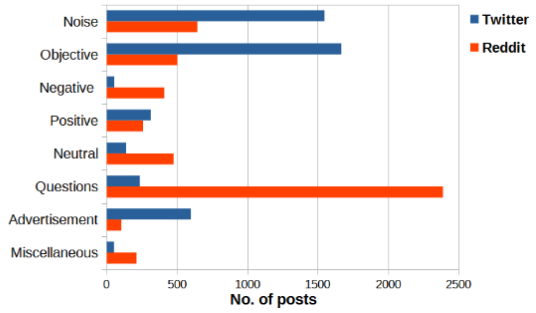

Figure 3: This the distribution of posts from both the media for each class

chy for labeling social media posts and observing potential patterns and rules. This initial sample informed the development of our labeling hierarchy (see **Figure 2**). Subsequently, a larger random sample of 4,500 posts was drawn from each platform (Twitter and Reddit) for manual annotation based on the established hierarchy. The annotators possessed strong English language proficiency, and familiarity with social media platforms and were independent of this study (not authors of the paper). Inter-annotator agreement was assessed using Cohen's kappa coefficient (McHugh, 2012). Both datasets achieved a kappa score close to 0.7, indicating substantial agreement. We further probed into the class-wise values and discovered that the annotators mostly had confusion in segregating between Questions and Miscellaneous and Neutral and Miscellaneous where the agreement was close to 0.6. It was relatively easier to detect noise where agreement was close to 0.9. For the rest, it was close to 0.7. Disagreements were resolved through mutual discussion, and ultimately, **only labels with high annotator agreement were retained in the final dataset (i.e. no datapoint is removed from the corpus)**. This resulting dataset of 5,500 data points constitutes the "Gold Standard Opinion Dataset" encompassing both social media platforms. The dataset is further divided into a training set ("Gold Standard Opinion Train" - 5,000 data points) and a testing set ("Gold Standard Opinion Test" - 500 data points). These are the two datasets used for training and evaluation for the experiments on opinion dataset for each of the social media platforms. The training data distribution is illustrated in **Figure 3**.

3.3 Creation of the Silver Standard dataset

The size of the dataset described in **Section 3.2** may be insufficient for training complex neural network models. To address this limitation, we present a technique to augment the annotated data via an automated approach. This approach results in the creation of the "Silver Standard Dataset" generated through a weakly supervised learning process.

Firstly, we trained a Support Vector Machine

Dataset	# Reddit posts	# Tweets
Gold Standard Train	5000	5000
Silver Standard (0.99)	19049	501

Table 3: Statistics of the annotated Reddit and U.S. Twitter posts. The confidence score for the Silver Standard is 0.99.

(SVM) classifier[3] on the 5,000 manually annotated posts ("Gold Standard Opinion Train" dataset). To extract text embeddings for the classification process, we utilized the TensorFlow Hub Universal Sentence Encoder (Google, 2020). Subsequently, the trained SVM classifier was applied to a larger set of unannotated posts. This process resulted in the classification of new instances into the predefined categories with associated confidence levels. We then selected instances classified with a confidence level exceeding 0.99, forming the "Silver Standard Training Dataset". This automated approach yielded 19,049 additional posts from Reddit and 501 posts from Twitter. **Table 3** summarizes the data statistics, including details on both the Gold Standard and Silver Standard datasets.

To leverage the expanded dataset, we retrained the SVM model by incorporating both the initial gold standard data and the newly acquired silver standard data. The model's performance was evaluated on a held-out test set containing 500 data points. As illustrated in **Figure 2**, the model trained with the combined dataset exhibited improved performance, particularly for levels 2 and 3 of the labeling hierarchy. This improvement is likely due to the increased training data volume, especially for these lower levels where the gold standard data becomes sparser.
The distribution of the silver standard dataset is provided in the Appendix Section 7.

3.4 Creation of QnA Dataset

Our gold-standard opinion dataset from Reddit has a high abundance of questions/queries largely because posts were taken from subreddits which are like communities on cryptocurrencies and the members of these communities often engage in question-answering activities on pertinent issues. Though not on such high frequency, question- answering activities are also observable on Twitter. The idea behind this is to make a dataset with questions/queries from the gold and silver standard datasets and answers from the comments of the posts. The average comment per question ratio is approximately 15 for Twitter and 30 for Reddit. For Twitter, each question was of 35 characters while for Reddit this number was 57 characters, on average. For Twitter, each answer was of 15 characters while for Reddit this number was 58 characters, on average

To assess the relevance of the comments, we used a scoring system depending on how much score is received by each of the comments. We have excluded the question posts that had fewer than 5 comments, as they can inflate the final scoring. The annotators first chose the top 30 to 50 comments for each of the posts and after manual observation of several random posts, it is seen that the top 10 comments have the highest relevance guarantee overall. Hence, the top ten comments with the highest score are checked manually and if found relevant are marked as relevant for the question, the rest are marked as non-relevant. On these top 10 posts, Cohen's kappa (McHugh, 2012) values were close to 0.9, and further disagreements were resolved through mutual discussion between the annotators. The statistics of the Question-answering dataset are shown in **Table 5**. **Table 4** shows some example questions followed by one relevant answer. We see that the questions are on the market stability of some specific cryptocurrencies or general queries regarding cryptocurrencies.

Question-Answering datasets		
Media	**Questions**	**Comments**
Reddit	946	29016
Twitter	427	4890

Table 5: **Statistics of the Question-Answering dataset**

4 Results and Analysis

In this section, a comprehensive description of the conducted experiments is provided. Predominantly, the experiments are divided into two main categories: (i) experiments utilizing the opinion dataset, and (ii) experiments utilizing the Question and Answer (Q&A) dataset. Each category encompasses a variety of experimental approaches. Additionally, an extensive analysis of the experiments conducted is included, offering insights into the findings and methodologies applied.

4.1 Experiments on the opinion dataset

To check the accuracy of the models trained on the proposed datasets, we designate **Gold standard**

[3]We employed the implementation available at https://scikit-learn.org/stable/modules/svm.html

Question(Twitter)	are you buying any #altcoins yet?i have been buying a few
Relevant Answer	buy #IOTA & Check us out man! We are an unfound gem. :P #bitdog#dogecoin
Question(Twitter)	which #crypto game releases are you most excited for?
Relevant Answer	Check us out man!#FLOKI upcoming metaverse game Valhalla.To be released end of year
Question(Reddit)	Can somebody help me understand why crypto is a hedge against inflation?
Relevant Answer	Because crypto's appreciation in value outpaces fiat's depreciation in value
Question(Reddit)	Another survey predicts $1m Bitcoin by 2030-am I right in thinking this is probably impossible?
Relevant Answer	Yup impossible that it will take that long bro

Table 4: Relevant answers of the questions based on likes/upvotes from the QnA dataset

Opinion Test datasets (separate for both Twitter and Reddit) that contain manually annotated 500 posts. The test sets are balanced i.e., contain a more or less equal number of posts across different classes. We experimented with classifying all the classes (8 classes) in the leaf nodes of the Tree 2. The experimentation is done in the following three ways:

1) Using **Universal Sentence Encoder** (Google, 2020) to generate vector representation of the texts and then use the classifier models to test accuracy on the test dataset, using **SMOTE** (Chawla et al., 2002) to deal with the class imbalance.

We run some standard classifiers like Support Vector Machine (SVM), Gaussian Naive Bayes (GNB), Random Forest (RF), Decision Tree (DT) and Multi Layer Perceptron (MLP)[4] where the models were trained on Gold Train Standard dataset combined; tested on Gold Test. The best results are mentioned in **Table 6**. The fine detail of the experiments is described in Appendix (**Section 8.2**)

Medium	SVM	RF	GNB	DT	MLP
Twitter	**75.29**	73.78	70.50	67.87	74.48
Reddit	**70.72**	59.47	66.89	57.34	69.52

Table 6: The F1 scores from the models with universal sentence encoder embedding (All the model parameters are optimized to get the highest F1 score)

2) Then we deployed BERT models particularly **BERTWEET** (Nguyen et al., 2020) for tweet dataset and **Vanilla BERT** (Lu et al., 2024) for Reddit texts (as the Reddit texts are long and Vanilla BERT works well with longer texts) to generate the embedding of the texts(for training the prior models) and then train the respective models for the classification task. Moreover, we have fine-tuned the respective BERT models for classification (each dataset on each of the models)The best results are mentioned in **Table 7**. The finer details of the experiments is described in Appendix (**Section 8.3**).

Medium	SVM	RF	GNB	DT	MLP	BERT
Twitter	75.26	76.93	70.52	74.29	78.69	**80.43**
Reddit	71.37	66.03	63.75	61.46	70.29	**73.73**

Table 7: The Macro F1 scores from the models with BERT embeddings(All the model parameters are optimized to get the highest F1 score)

3) This is a supervised fine-tuning setup. We fine-tuned three of the most popular LLMs, **GPT2** (Radford et al., 2019) (124 Million parameters), **GPT**3.5−**Turbo** (Brown et al., 2020) (174 Billion parameters), and **LLaMA2**[5] (7 billion parameters) (Touvron et al., 2023) for the classification task with our dataset with proper prompting 14 on page 12. The fine detail of the experiments is described in Appendix (**Section 8.4**). **Table 9** mentions the best prompting results.

Used Prompts: This is the fine-tune prompt that produced the best result across the models:

Classification Prompt:
You are a classifier that analyzes the emotion from a given post. A post can be classified into one of the following categories: noise, objective, positive, negative, neutral sentiment, question, advertisement, or miscellaneous.
Based on these categories, classify the post:
Example: {text}
Label: {Corresponding Label}

Table 8: Classification Prompt for fine-tuning LLMs on Opinion dataset

Medium	GPT2	GPT3.5	LLAMA2
Twitter	82.20	**88.00**	85.60
Reddit	74.16	**85.00**	81.67

Table 9: Macro F1 score from finetuning LLMs

[4] https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

[5] LLAMA2 and GPT2 are fine-tuned on 4 bit quantization using **QLORA**

4.2 Experiments on the QnA dataset

We address the Question-Answering problem on our dataset as a ranking tasks in both unsupervised and supervised setups.

Unsupervised: After determining the ground truth, we proceeded to calculate the **BM25** score (Robertson and Zaragoza, 2009), for each question and the comments associated with it. BM25, also called Okapi weighting, is a probabilistic Information Retrieval model for ranking documents for each query based on their term frequencies, document frequencies and document lengths. It has been used successfully in many ranking tasks. However, it is primarily based on term-based overlap between the query and the documents. In our task, we sorted the comments for each question in decreasing order based on their respective BM25 score. We evaluated the ranking using Mean Average Precision (Christopher D. Manning and Schütze, 2008) which measures the effectiveness of the ranking by considering the position in the ranking and the relevance of the comment. Note that, for this task, MAP is reported on the whole of Reddit and Twitter datasets. These results are shown in **Table 11** as Unsupervised (Whole).

Supervised: We have also done a supervised classification task on the question-answering dataset. It can be thought like a binary classification task with the dataset stating whether a question and comment pair is relevant or not. We have paired each question (post/tweet) with the comments/replies collected and marked the top-10 voted comments (also manually verified) with the question as correct (label 1) and the rest as incorrect (label 0). We experiment with three types of models: SVM with sentence embeddings, a finetuned BERT model variants, and finetuned generative language models. We split the two-class dataset in an 80-20 split and trained a classifier (SVM, as it produced consistent performance in opinion classification) for classification. We observed a 1:2 ratio in the labels; hence, we used smote (Chawla et al., 2002) to balance the data (as a performance drop of 8-10% was observed). For training purposes, we have concatenated each question embedding with comment/reply embedding obtained from the Universal Sentence Encoder (Google, 2020) embedding. For generating the results in a ranking setup, the probability score is calculated with each candidate's answer for each question in the test setup. Only the answers with a probability higher for class 1 (than 0) are chosen, and these answers are sorted in the decreasing order of these probability scores. Further, two BERT models (BERTweet on the Twitter dataset and BERT-Vanilla on the Reddit dataset) (Nguyen et al., 2020; Lu et al., 2024) are finetuned by simply joining the query text with the comment text in binary classification setup. Results are mentioned in **Table 11**

Finetuning of LLMs: This is also a supervised task where we finetune LLMs (GPT2, GPT3.5 and LLAMA2[6]) (Radford et al., 2019; Brown et al., 2020; Touvron et al., 2023) and make it a classification problem where we pair each question(posts) and answer (Comment/Reply) and fine-tune of 80% of the data with label either Relevant or not relevant and test it on rest of the 20%. The result of the experiment is put in **Table 11**.

Used Pompts: This is the prompt (See **Table 10**) that is used during fine-tuning on the QnA dataset.

QnA classification prompt:
Given a question-answer pair, it can be relevant or not relevant. You are a classifier that analyzes whether given a question-answer pair is Relevant or Not relevant. Based on this, put a label on the question-answer pair. **Question**:{question} **Answer**:{answer} **Label**:{label}

Table 10: QnA classification prompt for finetuning LLMs

Retrieval Augmented Generation (RAG): We have used RAG on our base LLAMA2 (7B) model. For this, the dataset is divided into 80-20 split. LLAMAIndex is used to vectorize the data, and choromaDB stores the vector embeddings. Our model searches through the vector database and generates based on the retrieved data(See **Figure 5**). The model is tested by matching the output with the original answer by LLAMA2 itself ((Gao et al., 2024)) and also used RAGAs (Es et al., 2023) as an evaluation metric for relevance of the answer. **Figure 5** pictorially shows RAG. See **Table 11**. The fine detail of the RAG experiment is described in Appendix (**Section 8.5**)

[6]LLAMA2 and GPT2 are fine-tuned on 4-bit quantization using **QLORA** on Google Colab using V100 GPU.

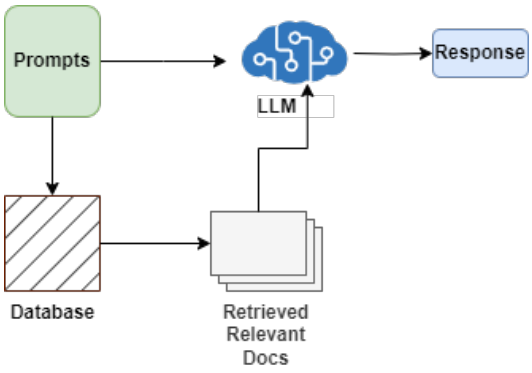

Figure 4: Figure shows the basic architecture of RAG

4.3 Analysis

Classification: The paper delves into a premise that is similar to sentiment analysis but with a different annotation scheme. Both the small foundation models as well as the large Language models performed notably well on the dataset. We believe the contributed dataset conveys the emotion of the users on the social media (see results: **Table 9, 7, 6**). Fine-tuned GPT-3.5 performed the best among all the models and it was able to detect minor text detection and change of labels, e.g. *#dogecoin merch* do not have any information about crypto if a semantic representation is considered, so the label is Noise which GPT-3.5 was able to detect is despite misleading superficial cues like *#dogecoin*. Noisy Tweets like *vote for lisa i vote for #mtvlakpoplisa* were plenty while the dataset was downloaded with crypto-related keywords and misclassified as positive by LLMs. This is the main motivation for the annotation scheme as most LLMs are trained on a general dataset fulfilling a sentiment analysis task for a general agenda that misses out on the usability of particular nuanced cases. So, the Noise class is one of the most significant classes and the models performed significantly well on detecting Noise after fine-tuning. There are some incidents when the best classification model GPT3.5 failed. There are some instances when the text is quite similar for both the labels, i.e consider the case: *Bitcoin Price Falls 40% As Coinbase Stock Drops 33% Cathie Wood Sees Bottom*. The actual annotated label is **Objective** as it shares some facts but it is a negative incident, the model predicted it to be **Negative**. Consider the text: *Bankrupt Crypto Lender Celsius Facing Federal Investigations* but here in this case the model was correct to predict it as **Objective**. There is some overlap between the positive class and the question class. Reddit texts are usually long and it is noticed that in some cases the user has put bulk information before stating the question. For those cases, many models have failed. Consider the text: *what is the cryptocurrency to you? when i first heard about cryptocurrency 4 years ago, i was very skeptical to say the least. but the more i look, the more i see that it is not a zero-sum game. it is something much bigger akin to the internet. to me, cryptocurrency is the new internet.* The actual label is **Question** but many models see a positive experience of the user hence predicting the label **Positive**. In most of the scenarios, the best model GPT-3.5 has performed significantly well over all other classes.

QnA: This paper introduces a new dataset of questions and answers (QnA) designed to train models for retrieving and generating information. This dataset is unique because a large portion of it consists of user queries about specific topics. This suggests that the QnA dataset could be valuable for creating systems that can effectively answer user questions.

We first tested the dataset using a retrieval method, BM25, but the results weren't very good. Next, we tried training models, both simple and complex ones, on question-and-answer pairs. However, these methods also didn't perform well, likely because the questions were very specific and because simply training models on existing data (zero-shot learning) wasn't enough. Later, BERT and LLMs were finetuned for the QnA task, and the results were improved significantly, particularly GPT-3.5. Further, we used a technique called Retrieval-Augmented Generation (RAG), which achieved significantly better results compared to other approaches.

The high variation in performance seen in **Table 11** might be due to the limitations of both BM25 and the simple model we used to understand the text (sentence encoding). Tweets are often short, which means they may not contain all the information needed to answer a question accurately. As shown in **Table 11**, models that have been trained on a wider range of data (pre-trained models) perform much better than models that rely solely on the QnA dataset itself.

This dataset is well-suited for use with modern tools like chatbots and AI assistants powered by large language models. This makes it a valuable resource in today's technological landscape. Additionally, incorporating user feedback on the generated answers into the training process (continuous

Medium	Unsupervised (Whole)	Unsupervised (Test)	SVM with Universal Sentence Encoder	SVM with BERT	GPT2	GPT3.5 Turbo	LLAMA2	RAG with LLAMA2	RAG with RAGAs
Tweet	**38.01**	20.99	15.99	**49.00**	42.00	**58.00**	54.00	**78.00**	74.00
Reddit	23.00	**32.05**	**49.50**	26.00	41.00	**51.00**	48.00	82.40	**84.20**

Table 11: The table shows the entire MAP score of the experiments done on the QnA dataset and the mean RAGAs score (in the last column). All the scores are scaled to the scale of 100.

learning) has the potential to further improve the performance of these systems.

5 Conclusion and Future Work

In this paper, we have developed a comprehensive corpus concerning cryptocurrency on a global scale, covering both Twitter and Reddit platforms. We observed that the categories of posts on these platforms differ both lexically and structurally. Consequently, we have systematically annotated the posts at three distinct levels. Additionally, user interest varies significantly across different media and countries. Reddit primarily serves to address queries, whereas Twitter hosts a diverse array of posts, including advertisements and factual content. Furthermore, we have compiled a question-answering dataset in this domain and evaluated the performance of several state-of-the-art (SOTA) question-answering methodologies. We believe that this dataset will prove invaluable in understanding user behavior and their attitudes toward emerging digital transaction mechanisms.

Furthermore, in subsequent phases of our research, we plan to employ Reinforcement Learning with Human Feedback (RLHF) (Zheng et al., 2023) and additional reinforcement techniques such as Direct Preference Optimization (DPO) (Rafailov et al., 2024) to enhance our models and impart explanability to the same.

Limitations

The main limitation of the dataset is that the dataset has not been updated to date due to restrictions in scraping data from the official APIs of Twitter and Reddit. However, this also ascribes to the value of the dataset, as to our knowledge it is the last dataset on cryptocurrency scraped from Reddit and Twitter. The size of the dataset is one of the limitations. We have managed to annotate 10k posts with our hierarchy from both the social medias. Firstly, initial funding (for manual annotation) and then restrictions from official APIs restricted us from scraping more data. In the later part of the experimentation, we have used high-end LLMs like GPT3.5, LLAMA2, GPT2 which are expensive, computationally as well as economically. Most of the experiments have been done in Google Colab with a Colab Pro+ subscription. So, running models is very expensive and causes carbon emissions. There are a lot of smaller models that require less consumption of energy and we should have used that before these heavy-weight models, with a lower accuracy consideration. As the dataset size is small we have shown one way to expand the dataset: the silver standard dataset. A lot of research could be done in expertise on the methodologies of dataset expansion. But as it is a dataset paper we have only focused on explaining our own dataset and usability without investigating methods for dataset expansion.

Ethical Considerations

We have duly subscribed for the paid LLMs (GPT) used in this paper. Also, the annotators have been compensated commensurate with their efforts.

Acknowledgement

This work was supported in part by, the Science and Engineering Research Board, Department of Science and Technology, Government of India (No. SRG/2022/001548). Koustav Rudra is a recipient of the DST-INSPIRE Faculty Fellowship ([DST/INSPIRE/04/2021/003055] in the year 2021 under Engineering Sciences).

Appendix

5.1 Human Annotators

Two human annotators are involved in annotating the dataset; both are quite well-versed in English and familiar with social media datasets. Each of them was given the definition of the classes and the initially sampled dataset of 1000 data points from one particular social media and was asked to annotate the rest of the 4500 dataset based on that. The annotators are paid as per standard rates for annotation.

5.2 Model Parameters

We have used 5 language models that require GPU to produce fast results. The model parameters are stated in the **Table 12**.

Model	Number of Parameters
BERTWEET-base	135 Million
Vanilla BERT	340 Million
GPT2	345 Million
GPT3.5	175 Billion
LLAMA2	7 Billion

Table 12: Parameters of the pre-trained models used

All the models (except GPT3.5) are loaded from hugging face with authentication at certain models. GPT2, LLAMA2 is fine-tuned in 4bits using QLORA and SFT trainer on V100 GPU. The BERT model are trained using *seq2seq* trainer on V100 GPU.

5.3 Experiment Platform and cost

All the codes are run on Google Colab with a subscription to Google Colab Pro+. For running the experimental setups, we have used almost 500-1000 computational units on Google Colab, charging approximately 100 US dollars. We now explain the hierarchy as below.

6 Creation of Hierarchy

Level 1

The initial goal is to differentiate between objective and subjective texts. Objective texts are common facts, price, news and subjective texts are opinions of people on cryptocurrency. During manual observation of a random initial sample of 1000. It is observed that a bunch of texts (almost 30%) that are not quite related to cryptocurrency or do not have any information about crypto but are still posted along the crypto hashtags. There are some texts that are incomplete, i.e., posted with a photo or a bunch of photos with hashtags only. We have put those in Noise class as we are only considering text. The rest of the dataset is manually annotated in two classes, namely **Subjective** and **Objective**.

Level 2

In this level, the initial goal is to segregate all the opinions (**Subjective** texts) into three major classes, namely **Positive**, **Negative** and **Neutral**. **Positive** texts are subjective opinions that have a positive attitude towards Cryptocurrency. Similarly, **Negative** texts are negative opinions towards Cryptocurrency. **Neutral** texts are neutral in nature, i.e., have neither positive nor negative attitudes.

Level 3

In the previous level, it is seen that a large number of posts($\approx 70\%$) are assigned to the class **Neutral**. Through manual observation, it is seen that the a bulk amount of Neutral texts are **Questions** (for the Reddit dataset) or **Advertisement** (for the Twitter dataset) .Hence, these two labels are created. Also, from manual observation, it is seen that some texts are neither **Question**nor **Advertisement** but pose neutral opinions. Hence, the label **Neutral sentiment** is created. In very rare cases, it has been seen that some texts sound like advertisements with over-promising gifts or sound like scams. These texts do not belong to any of our previously mentioned classes. Hence, we have created another label **Miscellaneous**.

See the **Table 2** where examples from different classes are stated.

7 Distribution of the Silver Standard Dataset

Table 13 shows the distribution of Silver standard dataset across both sub-corpora from Twitter and Reddit.

Label	Tweet	Reddit
Noise	79	5878
Objective	290	2125
Positive	35	596
Negative	7	1285
Neutral Sentiment	14	1041
Question	12	6568
Advertisement	62	1456
Miscellaneous	2	100
Total	501	19049

Table 13: The distribution of Silver standard dataset

8 Experiments

The experiments are designed to test the performances of the datasets created under several standard models and metrics. This shows the usability of the dataset w.r.t modern tools as well as the validity of the annotation scheme (Hierarchy) . The Experimental setup is divided in two major categories having subcategories within. In the later sections, all the experiments are noted in detail and further clarification of models and techniques are discussed in follow-up chapters.

8.1 Experiments on Opinion Dataset

This section is strictly devoted to the experimental set-up used on Opinion Dataset. This section is further divided into three major categories of experiments that have been done in this project. From low-level ML models to high-level deep learning models are used for the text/sentiment classification tasks performed on the **Opinion Dataset.** The **Gold Standard Train** (5000) and **Gold Standard Test** (500) are used with proper parameter tuning to train and test the models, respectively. The dataset is flattened to an 8-class dataset for training and testing, taking the leaf nodes of the **hierarchy 2**. This flattened dataset from the Opinion dataset is used in all of the classification tasks mentioned later.

8.2 Experiments on low-level ML Models

For analyzing the text through ML models, the text data has to be converted into numerical, more precisely vectors(embeddings) . There are already some pre-trained models that can produce vector embeddings, that can be fed to the ML models, and the model can analyze the text based on embeddings. Here the model, **Universal Sentence Encoder**[7] (by Google) is used to generate the embeddings. This model creates a vector of 512 dimensions for each text.
A set of low level models are used for text/sentiment classification task. These models are **Support Vector Machine (SVM)** , **Decision Tree (DT)** , **Random Forest (RF)** , **Gaussian Naive Bayes (GNB)** and **Multi Layer Perceptron (MLP)** . All the models are optimized to attain maximum performance w.r.t **F1 scores**.
Before feeding the texts to generate embeddings, a basic text processing pipeline has to be done. It is described below:

1. **Downcasing:** This means lowering down the English characters.

2. **URL Removal**: Removing all the possible links and URLs from the texts.

3. **Extra Space Removal:** Removing all the extra space from the texts.

4. **Removal of Emoji:** As we are only considering text inputs, we have removed all the emojis from the texts

After this preprocessing, the text is fed to the **Universal Sentence Encoder** model to generate embeddings. The model works in two ways; first it **tokenize**[8] and then generate the embedding (vector) depending on the token sequence.
These 512 dimension-vectors with appropriate labels are then fed into the models for training and testing after reducing the class imbalance with **MultiSMOTE**(Chawla et al., 2002).

8.3 Experiments on Encoder block of Transformer

Here, two of the pre-trained BERT models are used for sentiment analysis. These models are **BERTWEET** (for the Twitter dataset) and **BERT-Vanilla** (for the Reddit dataset) . First, we have done the experiments where we have generated embeddings from these BERT models and classify with the previously mentioned classifier models (See **Section 8.2**). These models are then finetuned with text data and labels. The models are fine-tuned with appropriate model parameters to attain maximum F1 scores.

8.4 Experiments on Decoder block of Transformer

Transformers' decoder block is designed to generate texts based on the previous context. Here, the decoder block is used to generate the label of a text coupled with the appropriate prompt. The GPT2 & 3.5 models and LLAMA2 are fine-tuned with appropriate prompts (See **Prompt 14)** and parameters to attain maximum F1 scores.

[7]The model is available in the following link: Tensorflowhub Universal Sentence Encoder

[8]**Tokenization:** It means chopping the original text in little chunks of texts(May or may not be meaningful semantically) .

Classification Prompt:
You are a classifier that analyzes the emotion from a given post. A post can be classified into one of the following categories: noise, objective, positive, negative, neutral sentiment, question, advertisement, or miscellaneous.
Based on these categories, classify the post:
Example: {text}
Label: {Corresponding Label}

Table 14: Classification Prompt for Fine-tuning LLMs on Opinion Dataset

8.5 Experiments on QA: RAG

We have performed **RAG** (Retrieval Augmented Generation) on our QA dataset. Retrieval finds relevant information from a vast text corpus, while generation uses that information to create coherent and relevant text. Here, RAG is implemented on LLAMA2 with LLAMAIndex to tokenize the texts, and ChromaDB is used to store the vector embeddings. The entire pipeline is designed on **Langchain** (Pandya and Holia, 2023). The retriever searches the relevant comments in the database based on a question and then presents them to the LLMs that generate the answer to that question (See **Figure 5**).

Used Prompt:

1) This is the prompt(See **Table 15**) used to generate answers from the given context.

RAG answer generation Prompt:
< |**system**| > Using the information contained in the context, give a comprehensive answer to the question. Respond only to the question asked, response should be concise and relevant to the question. Provide the number of the source document when relevant. If the answer cannot be deduced from the context, do not give an answer.< /s >
< |**user**| >
Context:
context
—
Now here is the question you need to answer.
Question: question
< /s >
< |**assistant**| >

Table 15: Answer generation prompt of RAG on LLAMA2

2)This is the evaluation prompt(see **Table 16**), used to see relevance of the answer to the question. We have considered rating$\geq$ 3 as true and rest false.

RAG evaluation prompt:
You will be given a question-answer pair. Your task is to provide a 'total rating' representing how useful this answer can be to machine learning developers building NLP applications with the Hugging Face ecosystem to the question. Give your answer on a scale of 1 to 5, where 1 means that the answer is not useful at all, and 5 means that the answer is extremely useful.
Provide your answer as follows:
Your Answer:::
Evaluation: (your rationale for the rating, as a text)
Total rating: (your rating, as a number between 1 and 5)
You MUST provide values for 'Evaluation:' and 'Total rating:' in your answer.
Now here is the question.
Question: question
Answer: Answer
Your Answer:::

Table 16: Evaluation prompt of RAG on LLAMA2

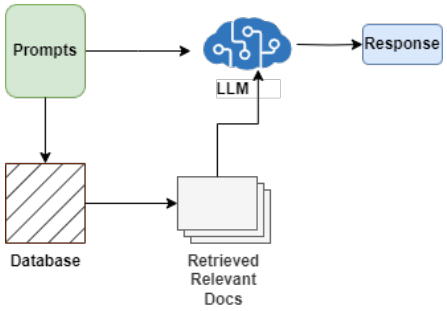

Figure 5: Shows a basic pipeline of RAG

For the evaluation of the RAG system, two techniques have been used. One is a metric **(RAGAs)** that evaluates how much the model hallucinates based on the query and retrieved comments. Another technique involves LLAMA2 as an evaluator that rates the answer based on the query and context (retrieved comments) on 1 to 5, and we only take the answer as valid if the rating is $\geq$ 3. Appropriate prompts are used to generate (see 15) and evaluate (see 16) responses from and on LLAMA2, respectively.

References

2023. Countries where bitcoin is legal and illegal.

2023. Profiling cryptocurrency influencers with few-shot learning (PAN at CLEF 2023): https://pan.webis.de/clef23/pan23-web/author-profiling.html.

Jethin Abraham, Danny W. Higdon, Johnny Nelson, and Juan Ibarra. 2018. Cryptocurrency price prediction using tweet volumes and sentiment analysis.

Igors Fridkins Ikram Ansari Mukul Vishwas Ali Raheman, Anton Kolonin. 2022. Social media sentiment analysis for cryptocurrency market prediction.

Fernando E Alvarez, David Argente, and Diana Van Patten. 2022. Are cryptocurrencies currencies? bitcoin as legal tender in el salvador. Working Paper 29968, National Bureau of Economic Research.

Dimosthenis Antypas and Jose Camacho-Collados. 2023. Robust hate speech detection in social media: A cross-dataset empirical evaluation. *Preprint*, arXiv:2307.01680.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. Smote: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

BC Cheong and K Lin. 2024. Crypto assets are property, specifically, choses in action, that are capable of being held on trust. *SAL Prac 2*.

Prabhakar Raghavan Christopher D. Manning and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. Ragas: Automated evaluation of retrieval augmented generation. *Preprint*, arXiv:2309.15217.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *CoRR*, abs/1802.00393.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Google. 2020.

Xiaojie Guo, Amir Alipour-Fanid, Lingfei Wu, Hemant Purohit, Xiang Chen, Kai Zeng, and Liang Zhao. 2019. Multi-stage deep classifier cascades for open world recognition. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 179–188, New York, NY, USA. Association for Computing Machinery.

https://docs.tweepy.org/en/stable/.

https://github.com/cvjena/twitter-flood-dataset. Https://github.com/cvjena/twitter-flood-dataset.

https://psaw.readthedocs.io/en/latest/.

Salud M. Jiménez Zafra, Giacomo Berardi, Andrea Esuli, Diego Marcheggiani, María Teresa Martín-Valdivia, and Alejandro Moreo Fernández. 2015. A multi-lingual annotated dataset for aspect-oriented opinion mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2533–2538, Lisbon, Portugal. Association for Computational Linguistics.

Aad van Moorsel Karolis Zilius, Tasos Spiliotopoulos. 2023. A dataset of coordinated cryptocurrency-related social media campaigns.

Xin Lu, Yanyan Zhao, and Bing Qin. 2024. Vanilla transformers are transfer capability teachers.

Omri Marian. 2013. Are cryptocurrencies 'super' tax havens?

Marry L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*, 22:276–282.

Johnnatan Messias, Vabuk Pahari, Balakrishnan Chandrasekaran, Krishna P Gummadi, and Patrick Loiseau. 2023. Dissecting bitcoin and ethereum transactions: On the lack of transaction contention and prioritization transparency in blockchains. *arXiv preprint arXiv:2302.06962*.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Sa'id Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, Alípio Jorge, and Pavel Brazdil. 2022. NaijaSenti: A Nigerian Twitter sentiment corpus for multilingual sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 590–602, Marseille, France. European Language Resources Association.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets.

Hoang H. Nguyen, Dmytro Bozhkov, Zahra Ahmadi, Nhat-Minh Nguyen, and Thanh-Nam Doan. 2022. Sochaindb: A database for storing and retrieving blockchain-powered social network data. In *Proc. ACM SIGIR*, page 3036–3045.

Keivalya Pandya and Mehfuza Holia. 2023. Automating customer service using langchain: Building custom open-source gpt chatbot for organizations. *Preprint*, arXiv:2310.05421.

Soham Poddar, Azlaan Mustafa Samad, Rajdeep Mukherjee, Niloy Ganguly, and Saptarshi Ghosh. 2022. Caves: A dataset to facilitate explainable classification and summarization of concerns towards

covid vaccines. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3154–3164, New York, NY, USA. Association for Computing Machinery.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Syed Shams Sanshao Peng, Catherine Prentice and Tapan Sarker. 2023. A systematic literature review on the determinants of cryptocurrency pricing. *China Accounting and Finance Review*.

Chahat Tandon, Sanjana Revankar, Hemant Palivela, and Sidharth Singh Parihar. 2021. How can we predict the impact of the social media messages on the value of cryptocurrency? insights from big data analytics. *International Journal of Information Management Data Insights*, 1(2):100035.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv e-prints*, arXiv:2307.09288.

Gaurav Verma, Niyati Chhaya, and Vishwa Vinay. 2020. "to target or not to target": Identification and analysis of abusive text using ensemble of classifiers. *CoRR*, abs/2006.03256.

Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, et al. 2023. Secrets of rlhf in large language models part i: Ppo. *arXiv preprint arXiv:2307.04964*.