Mohammad Maaz
Natural Language Processing for Data Science
Individual Final Report

Impact of Emotion on Bitcoin Price

2nd May 2022

INTRODUCTION

In our project, we aim to analyze the impact of emotion of cryptocurrency market players on the price of Bitcoin using the sentiment classification power of BERT. There were three phases in this endeavor.

1. Developing a dataset of Reddit comments from r/cryptocurrency
2. Training a classifier to predict the emotion exhibited in a text
3. Classifying the emotion in Reddit comments and comparing with the price of Bitcoin

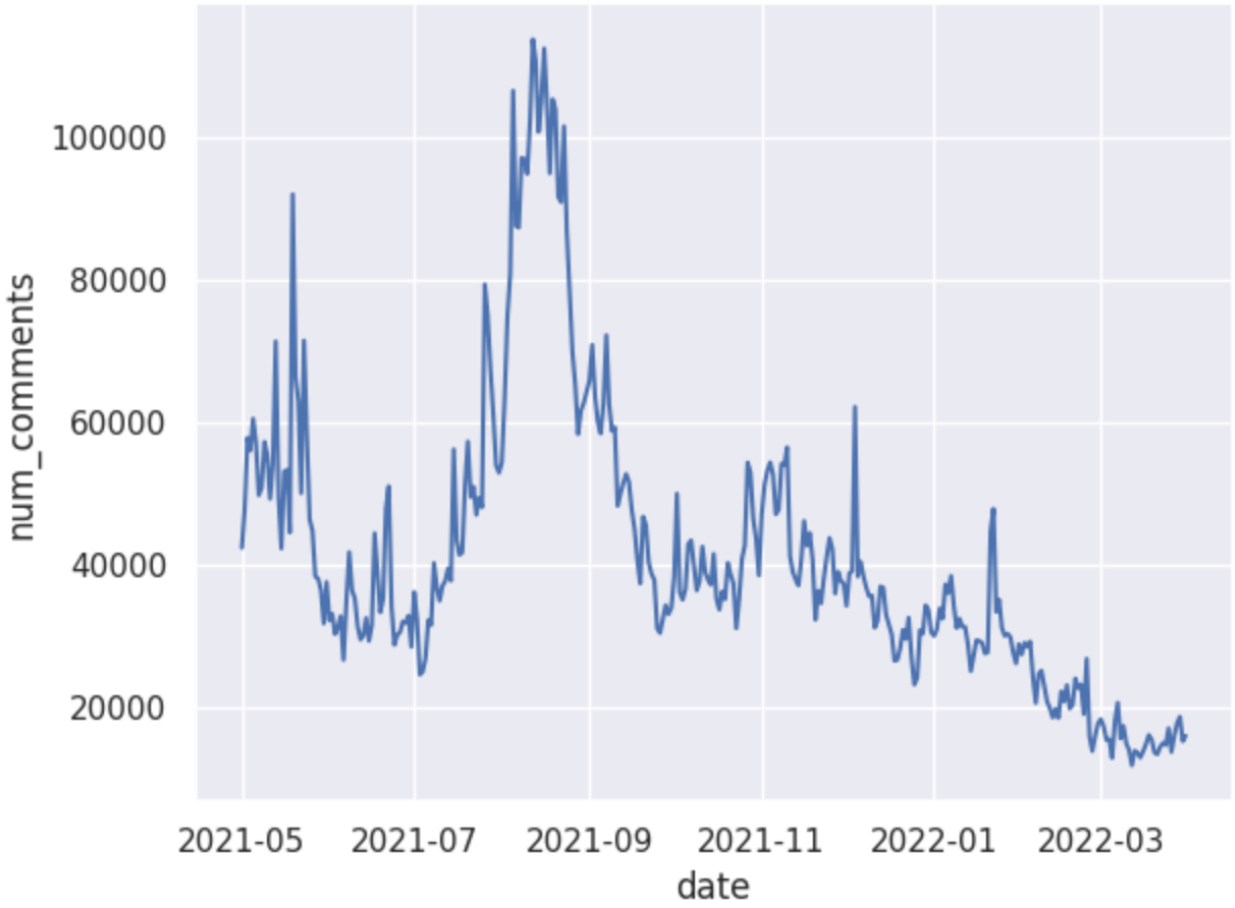
INDIVIDUAL WORK

My individual work in this project are the first and third phases mentioned above.

Dataset Development

I obtain Reddit comments using PMAW[1], a multithreaded wrapper for PushShift API[2]. PushShift is a RESTful API that gives full functionality of searching Reddit data. I am interested in comments, so I used the /reddit/search/comment endpoint. This API has a requests per minute limit of 60 and data points per request limit of 100, which is why I use PMAW to work on this in a multithreaded way to speed it up. Using this approach, we pulled 18 million comments from the cryptocurrency subreddit from 1st May 2021 to 1st April 2022. The runtime for this script was around a week.

[1] https://github.com/mattpodolak/pmaw
[2] https://github.com/pushshift/api



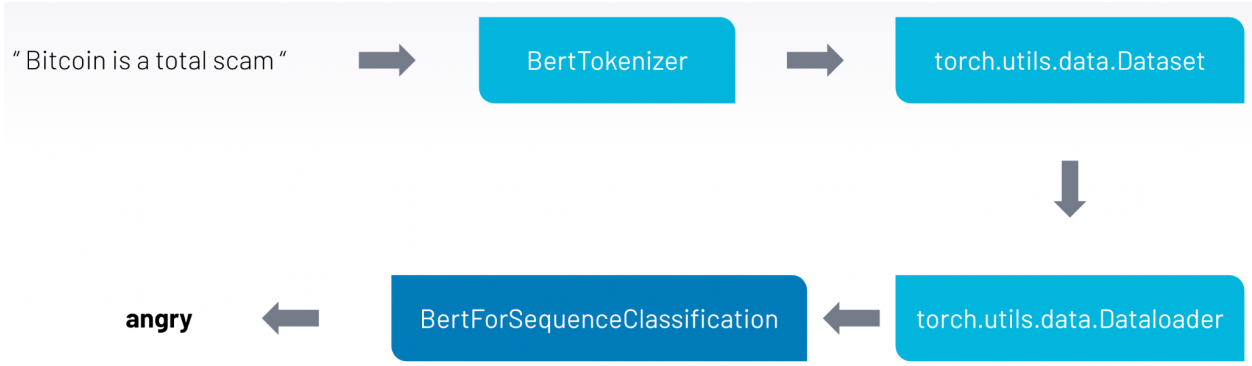
I obtain the historical Bitcoin price using the Yahoo Finance API, wrapped in the pandas_datareader library. I save only the adjusted close price, because that is a good measure for the price on a particular day.

The emotion[3] dataset was obtained from Hugging Face. This is a dataset of Twitter messages with six different emotions: anger, fear, joy, love, sadness, and surprise.

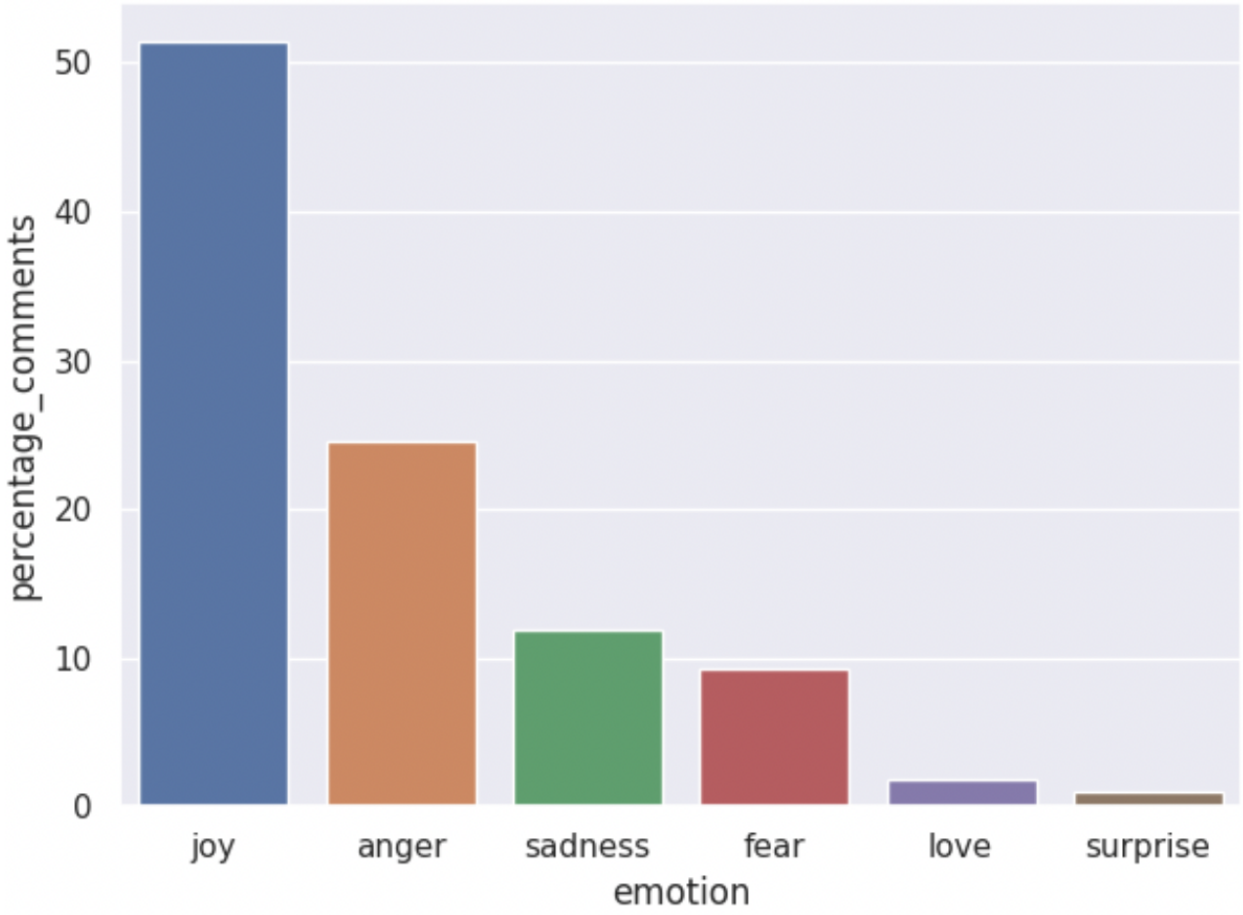
Emotion Prediction

I load the model trained by my groupmate from the last checkpoint. I then load a random subset of the 500k Reddit comments from our dataset of 18 million, tokenize each comment, and convert the smaller dataset into an object of a child class of the PyTorch Dataset class. I then define a PyTorch Dataloader object using this dataset, along with a collate function that pads the tokens to the same size. I then iterate across the Dataloader object and run the model on every batch. I then use the model outputs to get the predicted labels by using the argmax function and the id2label dictionary I created. This is outlined as follows.

[3] https://huggingface.co/datasets/emotion



The distribution of the predicted emotions is as follows.



RESULTS

With the 500k comments along with their UTC time and predicted emotion, I transform and pivot the data to get the percentage of comments that exhibited a particular emotion on a particular date. I then plot this along with the price of Bitcoin to observe trends.

We can see that sadness saw an all time low during the Bitcoin price increase of September 2021. However, it has increased since then, despite the new all-time high of late December. We can also see that joy is very highly correlated to the Bitcoin price. An interesting thing to note is that it lags price increases. There was an uptick in joyous comments as early as July 2021 even though the price boom happened in September of the same year. Joyous comments started decreasing in September 2021 and the crash of late 2021 happened mere months later. Angry comments are very interesting because they are inversely correlated with the price drops. This is expected because people tend to take out their anger on the same forum that encouraged them

to invest. This is indicated by the fact that crests of the graph of price and the trough of the graph of anger align.

CONCLUSION

The predictive power of transformers is evident in that they perform well in zero shot classification tasks. In our project, we used this very attribute to predict the emotion of Reddit comments. The correlation between the percentage of comments exhibiting certain emotions and the price of Bitcoin indicates that these predictions were reliable. Furthermore, our work also shows that cryptocurrency market analysts can harness NLP techniques to advise investing, because cryptocurrency prices are more volatile than traditional stock prices.

There are several possible improvements and ideas for future work possible in this project.

- Annotate the comments dataset in order to train a classifier on it
- Analyse the price of multiple cryptocurrencies instead of just Bitcoin
- Account for other events in the crypto space that might influence sentiment
- Predict on the entire dataset rather than just a subset

Percentage of Copied Code

No code was copied from the internet apart from documentations of the various libraries and previous coursework.

File	LOC	Copied%
bert_infer.py	67	0
preprocessing.py	37	0
prices_pull.py	14	0
reddit_pull.py	16	0
results.py	58	0

References

All references have been provided as footnotes.