Project 2 Presentation

Sentiment Analysis on Crypto tweets during Ukraine war

Team 3

Shivangi Gupta
Jaime Aranda
Peter Morales
David Lopez

Project Goals

https://github.com/sububer/sentimental

01 Use Python to obtain raw data from twitter and process it to make it readable for sentiment models

02 Use and compare machine learning model to predict price with sentiment and tweet metadata

03 Analysis of Twitter Data/ tweets based on Russia Ukraine conflict and crypto queries

Terms To Know

Sentiment Analysis

Sentiment analysis (or opinion mining) is a natural language processing (NLP) technique used to determine whether data is positive, negative or neutral.

Tokenization

Tokenization is essentially splitting a phrase, sentence, paragraph, or an entire text document into smaller units, such as individual words .
E.g. word_tokenize() , sent_tokenize()

Lemmatization

Lemmatization is the process of converting a word to its base form (lemma) considering the context and morphological analysis of word.

Query Detail

https://developer.twitter.com/en/docs/twitter-api/tweets/search/api-reference/get-tweets-search-recent

#query = "((#Bitcoin OR #BTC OR BTC OR #Bitcoin OR #HarmonyOne OR ETH OR #Ethereum OR DOT OR #Polkadot OR #cryptocurrency)(((#Ukraine OR #Russia)(war OR conflict OR crisis OR economy)) OR (#UkraineRussiaWar OR #RussiaUkraineWar))) lang:en -is:retweet"

query = "((#Bitcoin OR #BTC) (#Ukraine OR #Russia)) lang:en -is:retweet"

#fields=author_id,created_at,entities,geo,id,in_reply_to_user_id,possibly_sensitive,public_metrics

Sourcing and Processing Data

- **Twitter API Academic Research**
- **Retrieves up to 10mill/month**
- **1 project only**

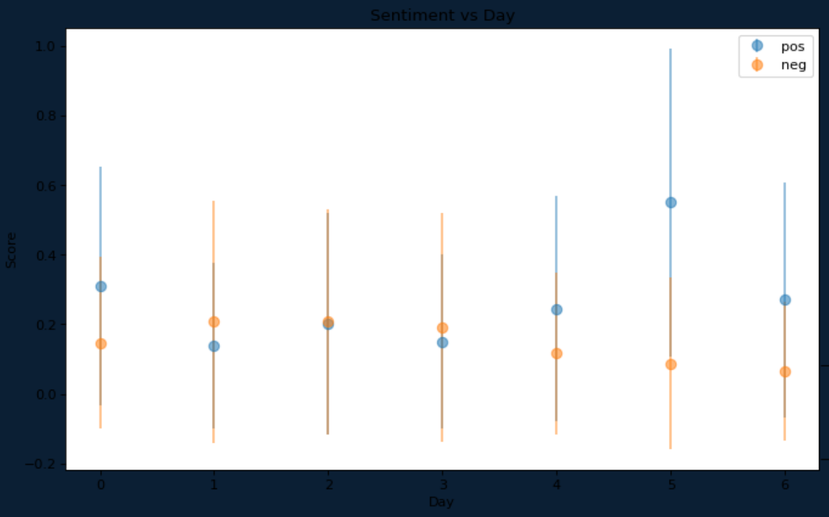
Steps for data processing:

- Case normalization/ standardizing text
- Removing Unicode Characters (Punctuation, Emoji’s, URL’s and @’s)
- Removing hyperlinks, marks and styles
- Removing Stopwords (words that don't value)
- Stemming and Lemmatizing text
- Tokenize tweets text

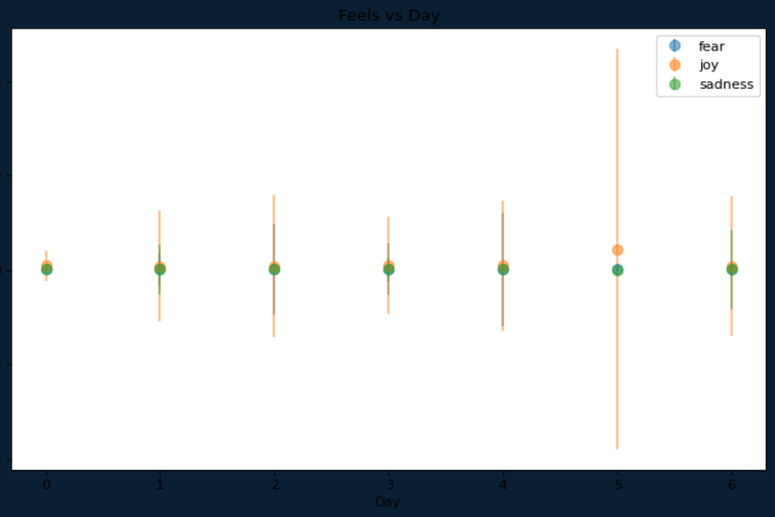
Models

Sentiment	Price Regression	Deep Price Regression
Pretrained Transformer SemEval 40k Tweets Predicts - Pos/Neg/Neu	Linear Regression Sentiment vs % Price Change Lasso regression : High Colinearity of tweet metadata features, reduce variance	Simple MLP Model Sentiment vs % Change Drop-Out Regularization

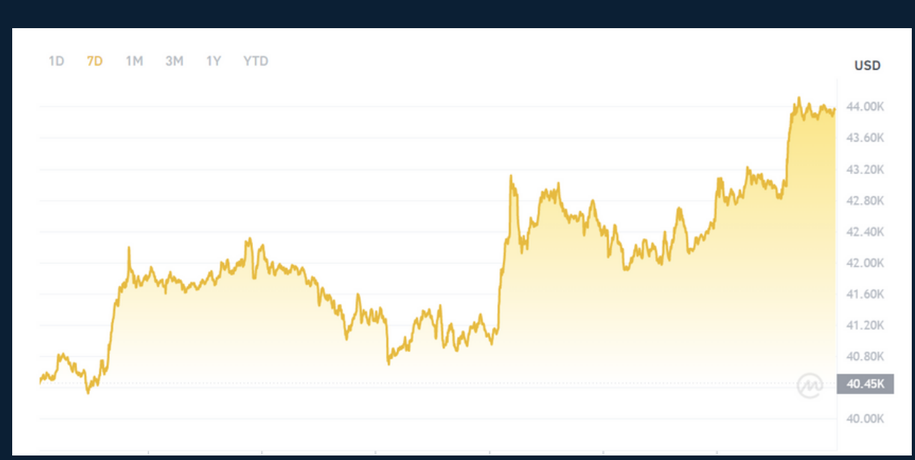
Visualizations



Sentiment



Emotions



Price of BTC Chart

Challenges

- Authoring the API query with the filters and operators for getting right Twitter data with minimum noise and pertaining to the required features and period for the analysis

- Preprocessing the unstructured tweet data to feed the model

Next Steps

- Dataset builder
  - Build out sentiment + tweet dataset
  - Train MLP & Linear Regressor to predict price
- Smart Contract for automated trading