Bitcoin Price Influence based on Sentiment Analysis of Relevant Tweets

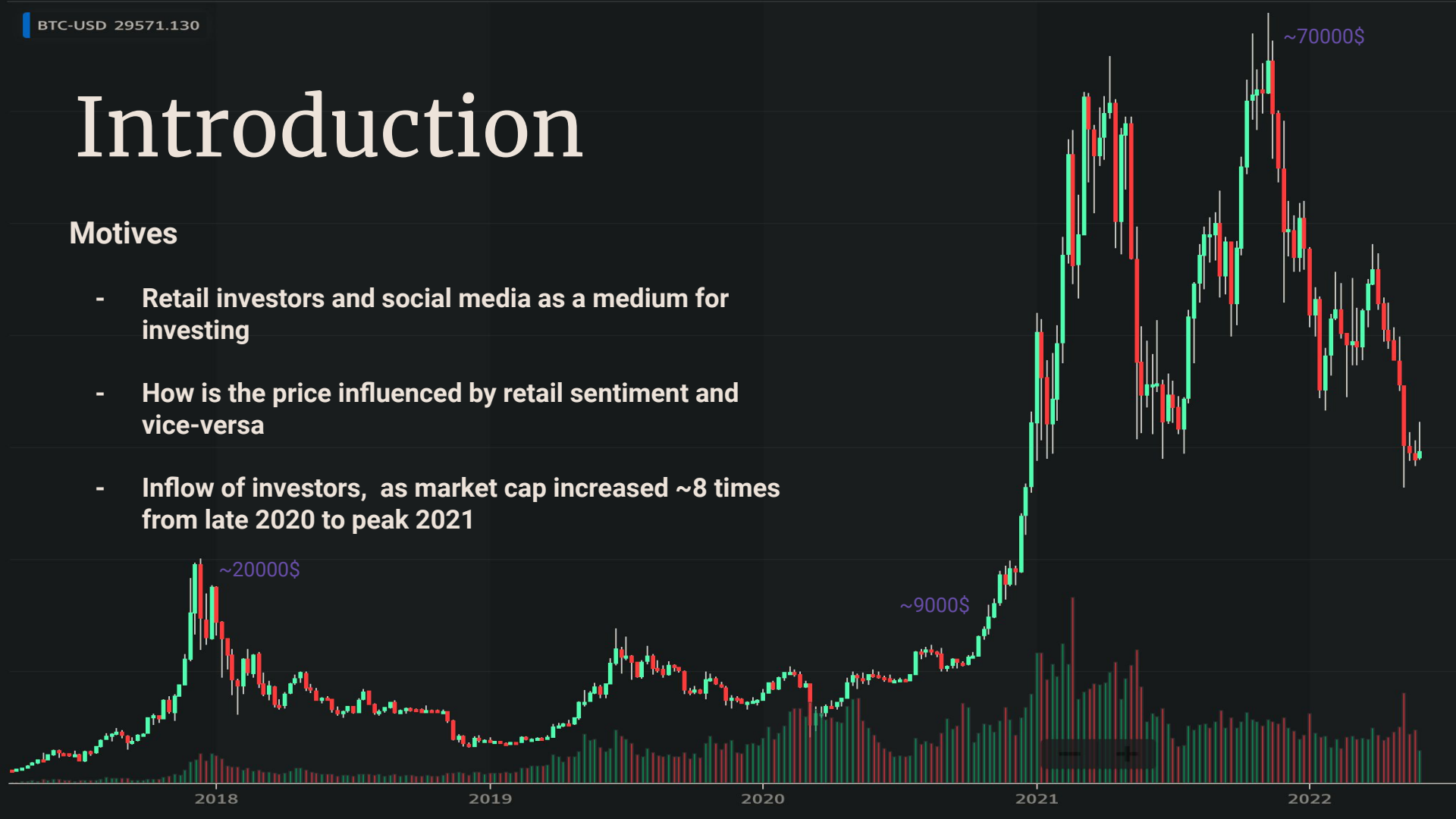
J. Brook, I. Postavaru, P. Ngare

Contents

Introduction

Motives

- Retail investors and social media as a medium for investing
- How is the price influenced by retail sentiment and vice-versa
- Inflow of investors, as market cap increased ~8 times from late 2020 to peak 2021



Datasets

Training dataset

- Kaggle: *Bitcoin tweets sentiment analysis*
  - 1.5 million BTC tweets from 2021
  - Sentiment tagged with VADER Sentiment
  - Used to create our own sentiment-tagging model

	date	tweets	score
0	2021-02-05 10:52:04	AT_USER AT_USER AT_USER right here w/ AT_USER ...	0.0000
1	2021-02-05 10:52:04	AT_USER AT_USER please donate bitcoin19 donate...	0.6597
2	2021-02-05 10:52:06	$sos market cap is 308 million. if they're min...	0.0000
3	2021-02-05 10:52:07	bitcoin btc current price (gbp): £34,880 like ...	0.3612
4	2021-02-05 10:52:26	AT_USER right here w/ AT_USER URL referral cod...	0.0000

(1519555, 3)

- API data
  - 500,000 tweets from 17/05/2022 - 23/05/2022
  - Collected using Twitter's Developer API

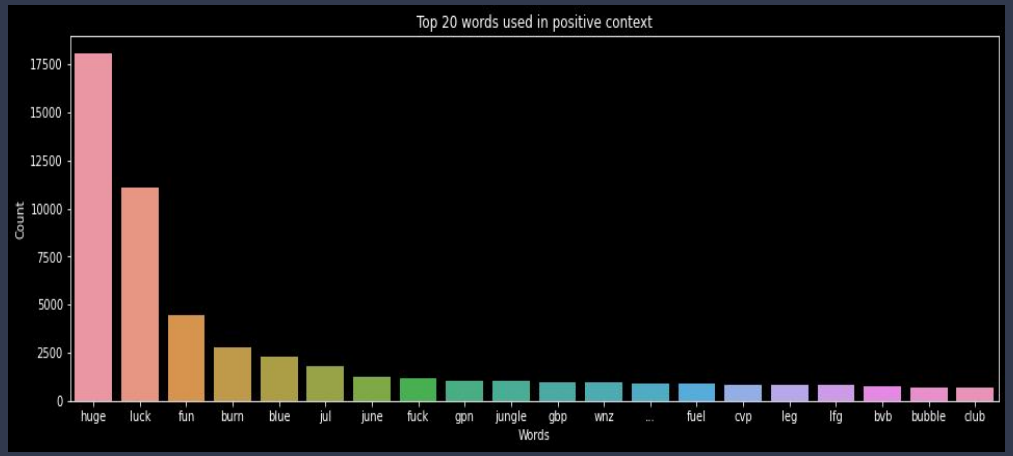
	Tweet	Date
0	current stats of delegatedonthate block find p...	2022-05-19 23:59:59+00:00
1	bbcworld for all those who be new to this work...	2022-05-19 23:59:58+00:00
2	smilingpunks floor price no gas fee polygon b...	2022-05-19 23:59:56+00:00
3	i be claim my free lightning sat from bitcoine...	2022-05-19 23:59:56+00:00
4	washingtonpost for all those who be new to thi...	2022-05-19 23:59:54+00:00

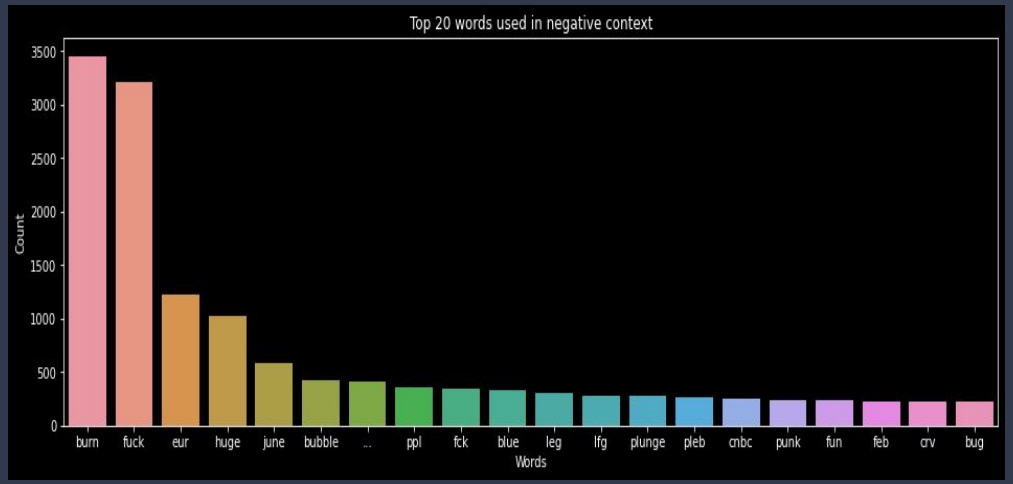
(500000, 2)

In-depth Kaggle dataset properties

Vader Sentiment

- "gold-standard" lexicon
- Tailored for social media content
- Lexical features such as acronyms, initialisms and slang with sentiment value
- ~40% of training data comes from tweets
- Sensitive to both the polarity and the intensity of a sentence (





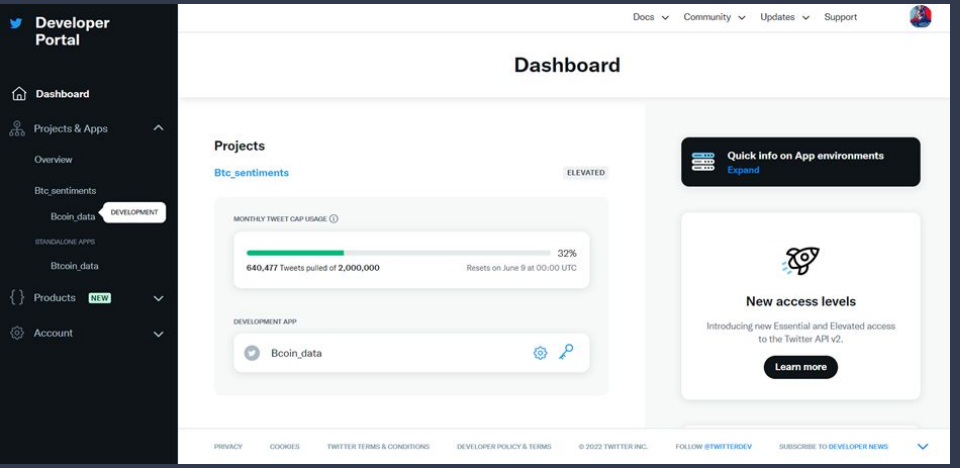
2018 Dataset

- ~ 50k tweets
- 0.0079 MSE on native test
- 0.1035 MSE on interpolated datasets

2021 Dataset

- ~ 1.5m tweets
- 0.0411 MSE on native test
- Cannot perform interpolated testing because of limited vocabulary of the training set
- 0.0606 MSE on native test with sliced training

API web crawling



- The Twitter API can be used to retrieve and analyze Twitter data.
- We crawled 500,000 tweets from the period between the 17th of May to the 24th of May 2022.
- We selected Tweets that contained any of the following words:
  - Bitcoin
  - bitcoin
  - Btc
  - btc
  - #Bitcoin
  - #bitcoin
  - #Btc
  - #btc

Preprocessing

Filtering

- Stop words, punctuation, links, @s

RegExp Tokenizer

- Highly customizable

NLTK POS-tagging

- Solid choice out of the box

WordNet Lemmatizer

- With WN converted POS-tag

Model Training Specifications

TF-IDF (term frequency inverse document frequency) is used as our **baseline** for word representation.

Term frequency **TF** (number of times a term appears in a document) multiplied by Inverse Document Frequency **IDF** (log of the number of documents over the document frequency of term).

Regression model, as we are trying to predict sentiment on a **scale** of -1 to 1.

Ridge Regression from SKLearn excels for this type of regression

Estimates coefficients of **multiple-regression models** in scenarios where **linearly independent variables** are highly correlated.

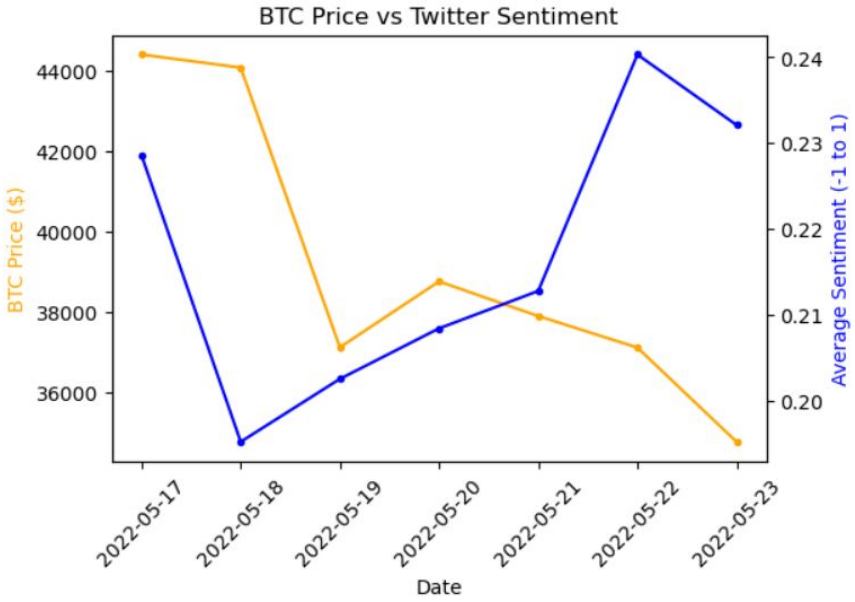
In this model, the **loss** function is a ***linear least squares*** and **regularization** is given by the ***l2-norm***

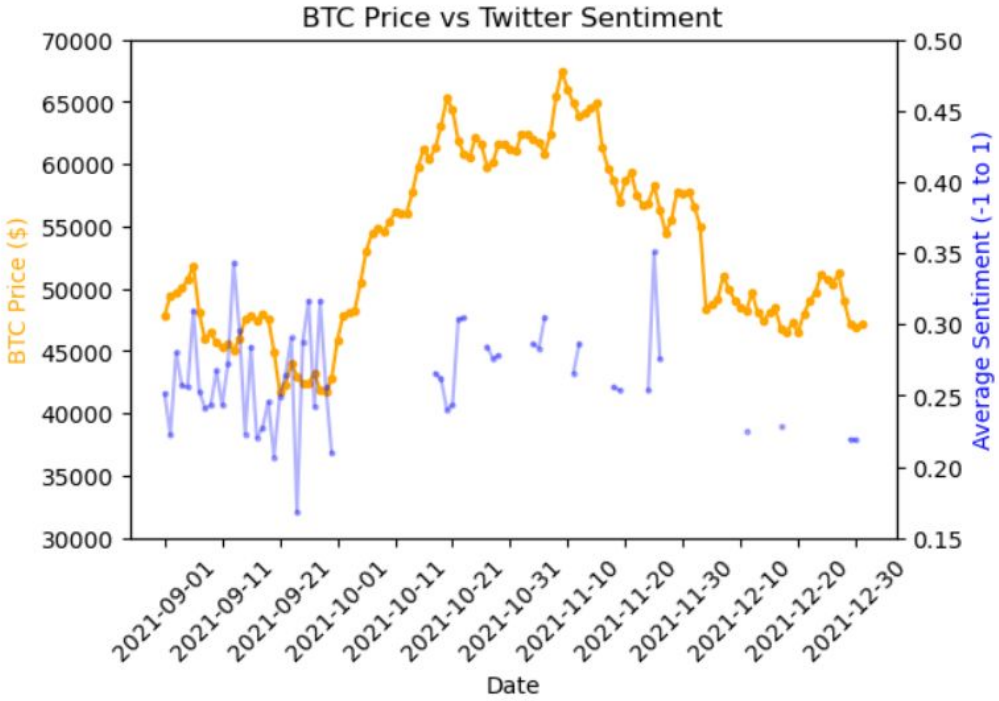
Bert-Model

- We also decide to try train a model based on BERT (Bidirectional Encoders Representations from Transformers) which is both a contextual and bidirectional language model.

Methodology

- We did text preprocessing (special tokens, padding, and attention masks) and built a Sentiment Classifier using BERT.
- The model was supposed to be Trained on the 1.5 Million Tweets Dataset:
  - Our features **X** are sequences of **tweets**.
  - Our target **y** is: the **Sentiment** tag for each Tweet.
- Finally, our model would be tested on the Tweets we Web Crawled through the Twitter Api and Evaluated by Accuracy score.





Graphs - 7 day vs 3 month

References

Abdali, S., & Hoskins, B. (2021). *Twitter Sentiment Analysis for Bitcoin Price Prediction*. Stanford CS229. http://cs229.stanford.edu/proj2021spr/report2/81988764.pdf

Pano, T., & Kashef, R. (2020). A Complete VADER-Based Sentiment Analysis of Bitcoin (BTC) Tweets during the Era of COVID-19. *Big Data and Cognitive Computing, 4*(4). https://doi.org/10.3390/bdcc4040033

Alexandrayuliu. (2021). *Bitcoin tweets sentiment analysis dataset*. Kaggle. https://www.kaggle.com/code/alexandrayuliu/bitcoin-tweets-sentiment-analysis/data