**Notes for dissertation:**

Kaggle links of interest:

**Comments on reddit dataset**

<https://www.kaggle.com/nickreinerink/reddit-rcryptocurrency/code>

**Headlines Dataset:**

<https://www.kaggle.com/kashnitsky/news-about-major-cryptocurrencies-20132018-40k?select=crypto_news_parsed_2013-2017_train.csv>

**Crypto Pairs, 1 minute resolutions from bitfinex (bitfinex = possibly questionable data, tether = probe in price manipulation, if manipulated, what does it mean for predictive power of data?):**

<https://www.kaggle.com/tencars/392-crypto-currency-pairs-at-minute-resolution/code>

Tutorial on how to get updates from bitfinex and append to historical data

<https://www.kaggle.com/tencars/getting-up-to-date-bitcoin-data>

**Ideas:**

Should probably focus on price prediction from text analysis as:

1. Easier to see the correctness of my work. Does my current prediction x match price y at time t. Focusing on fraud may take longer than 3 months of work, August 20th is not that far away, must be realistic.

2. Can I realistically claim fraud on a failed project? Lack of success does not equate to fraud, I could be measuring failure and labelling it fraud.

**Interesting papers:**

<https://advances.sciencemag.org/content/advances/6/51/eabd2204.full.pdf>

<https://royalsocietypublishing.org/doi/pdf/10.1098/rsos.170623>

On the efficient market hypothesis & bitcoin

<https://www.sciencedirect.com/science/article/pii/S0165176516303640?casa_token=mcdYzxemVkMAAAAA:lSNj38pAB6n-HAtb9YlM4a4Le6rR2vBXwcuzTXXqqdlYWVoKLYYZW0rcTc_s8KLc_U5hlrRr5Ig#b14>

1. Comparing bitcoin as an asset to gold is fine, stating it has higher volatility is questionable, gold prices heavily rely on futures market and are controlled that way, futures on bitcoin are at best in their infancy. Paper trading of gold (but more silver) heavily influences the price, hard to say how many “paper” ounces exist in comparison to real physical ounces. We can know exactly how many bitcoin have been created and the process is far more transparent than creation of paper gold and silver. See JP Morgan silver manipulation fraud:<https://www.cnbc.com/2020/09/29/jp-morgan-settles-spoofing-lawsuit-alleging-fraud-in-metals-trades.html>

One asset is older than human civilisation the other is younger than windows vista.

2. Results are that historical data suggests bitcoin is not weakly efficient. Arbitrage opportunities that people probably use.

<https://spia.princeton.edu/system/files/research/documents/Felten_SoK.pdf>

interesting overview, with interesting closing remark “future researches should attempt to measure volume of this bubble addressing to question “are we in the peak of Bitcoin bubble?””:

<https://poseidon01.ssrn.com/delivery.php?ID=111089002009102075004109070123091087034086041036045026092113020030105096098080114031022029052037057008050108097098098106103080122004033060060127093076082117076086113061041033106016085003070008027092099024108090018030009103028001122109095092110024092022&EXT=pdf&INDEX=TRUE>

<https://www.sciencedirect.com/science/article/abs/pii/S0167268120300676?casa_token=2zL8WzfPRUMAAAAA:sRBkdLHaqLyrH-4TOv2FQcvZyCaHfkuuprOwkrDN0Zki1zgTDVZhHI3bWqsD-vxL9SxALiPGYQI>

<https://www.sciencedirect.com/science/article/abs/pii/S0378437116309384?casa_token=2v52Y05u6lsAAAAA:uQ4BJVv2ziepRRiC9XJzNL-aBDjUKcFjGa95EObjZISK8aCs-ZixI0JlYQ1_HHXeI3MyLVDQ3LI>

<https://ieeexplore.ieee.org/abstract/document/8999092?casa_token=dwHaPNvqUfgAAAAA:vvmTQKmyHHCOBWlGhefY_08gG1HeCvrXTOc47C3mD_h796wSN653EDrMT9zHTB11M55dS0zF2bg>

<https://scholar.smu.edu/datasciencereview/vol1/iss3/1/>

<https://link.springer.com/article/10.1007/s10479-020-03575-y>

<https://www.frontiersin.org/articles/10.3389/fphy.2019.00098/full>

<https://media.neliti.com/media/publications/237311-autoregressive-integrated-moving-average-c8997cc8.pdf>

<https://www.sciencedirect.com/science/article/pii/S2214635019301534?casa_token=3fFzy6CR-FcAAAAA:qkCnuB-3K7DyWrX0WKtWEUqSGw2B8hSrr7E6ZsBgeQ9nrJmcHugJj9XJ8rzj7Bjc62YiYJzOxcw>

<https://onlinelibrary.wiley.com/doi/full/10.1111/manc.12352?casa_token=qO73xAarsLgAAAAA%3AsPh6WRbxMEO8VlQBVKLW4543S1HTTg5mROQ9KM-94bC8BjvBE6gKIBR6YmzXkAX59eEnSA1tId10MqTg>

<https://arxiv.org/pdf/1907.00558.pdf>

<https://storage.googleapis.com/plos-corpus-prod/10.1371/journal.pone.0161197/1/pone.0161197.pdf?X-Goog-Algorithm=GOOG4-RSA-SHA256&X-Goog-Credential=wombat-sa%40plos-prod.iam.gserviceaccount.com%2F20210322%2Fauto%2Fstorage%2Fgoog4_request&X-Goog-Date=20210322T150836Z&X-Goog-Expires=3600&X-Goog-SignedHeaders=host&X-Goog-Signature=>

Interesting way of doing sentiment analysis

<https://www.augmento.ai/wp-content/uploads/2019/06/Sentiments-and-topics-measured-by-Augmento-s-AI.jpg>

<https://www.augmento.ai/bitcoin-sentiment/>

#############

The mask slipping, when you are anonymous on the internet you truly behave as your unfiltered mind. Usually if you treat a thought as a bouncing ball, before you speak it you bounce this “idea-ball” against a filter, and it either:

1. Goes through the filter and gets spoken

2. Bounces back from the filter and you begin to reframe the idea so it passes the filter

The proportion of ideas spoken in person is very small:

(ideas spoken)/(total ideas) : ideas spoken – small, total ideas – big

On an anonymous forum where there is less social pressure and you do not feel you will face repercussions from badly framed ideas therefore we get a larger proportion of ideas spoken/written. Ideas on the internet are cheaper in a variety of social ways (less repercussions, people who judge are not close to you cannot impact your life)

(internet ideas spoken/(total ideas) – internet ideas spoken > ideas spoken | total ideas = same.

Numerator is larger. We see more of what drives humans/investors through the internet than in real life due to the lack of this “brain social filter”.

##########

Expand the lexicon? There is no way that VADER, being as general purpose as it is, truly captures all the emotion/sentiment in the current state.

<https://coinmarketcap.com/alexandria/glossary> , https://www.wallstreetbets.shop/blogs/news/dissecting-the-unique-lingo-and-terminology-used-in-the-subreddit-r-wallstreetbets

<https://stackoverflow.com/questions/40481348/is-it-possible-to-edit-nltks-vader-sentiment-lexicon>

FOR LEXICON FILE:

In vader, the class SentimentIntensityAnalyzer: loads this lexicon:

<https://github.com/cjhutto/vaderSentiment/blob/master/vaderSentiment/vader_lexicon.txt>

could potentially change directory to my own version of this with appended “crypto world phrases” and the cross over terms from r/wallstreetbets i.e. all the stuff below, maybe capture more sentiment BUT have to think to not introduce huge bias from my own personal opinion.

<https://www.researchgate.net/publication/324957692_Financial_Sentiment_Lexicon_Analysis>

<https://arxiv.org/pdf/1307.5336.pdf>