Analysing crypto news sentiment to predict bitcoin prices

Abhishek Kumar
Worldquant University

Abstract

The purpose of this paper is to analyse the ability of news content to predict cryptocurrency markets. The role of news announcement is central to pricing and revisions in pricing of any asset. Right from the 16th century news of ships arriving at ports with tradable goods resulted in the fluctuation of local market prices. In the digital age where TV screen flashes breaking news at microsecond frequency, the influence on the prices has never been more profound. Using news articles to predict markets has always been a bottleneck for traders, the major issue being transforming words and semantics to financial numbers. The meteoric rise of natural language processing in other fields has finally made this task possible for humans. Some common ways include transforming raw texts to bags of words, one hot encoding or advanced word embeddings and feeding them to ML models which is attempted in this paper.

Keywords: *NLP, Sentiment Analysis, Machine Learning, Crypto, Bitcoin.*

INTRODUCTION

Financial firms use a wide variety of methods to predict asset prices. Developing and testing economic theories, modelling prices using complex statistical models or using alternate sources of data, traders are always in search of even the smallest Alpha which can add improvement to their existing model and make them more money. The alternate datasets include research reports, microeconomic variables, microstructural data and social media and news articles data.

In this paper we will explore one such alternate data for alpha generation, the news articles. Specifically, we judge the predictive ability of news articles published in leading cryptocurrency blogs for prediction of cryptocurrency

markets, especially bitcoin. Bitcoin was chosen as a representative of the entire crypto-currency market. This was done due to the fact that it has the highest market cap and the highest traded volume across most of the existence of the cryptocurrency universe.

LITERATURE REVIEW

There is a vast literature on application of sentiment analysis on equity markets. Many authors have attempted to classify the sentiment of reported corporate filings at SEC. Initial works in this field focused on using the count of positive and negative words in the article and classifying documents accordingly. In one of the earlier works, *Tetlock(2007)* in the paper titled *Giving Content to Investor Sentiment: The Role of Media in the Stock Market* counts the positive and negative sentiment words in a famous wall street journal column and successfully predicts the broader markets. They used a prebuilt classified Harvard word dictionary to accomplish the task.

In another paper titled *More Than Words: Quantifying Language to Measure Firms' Fundamentals TETLOCK, TSECHANSKY, and MACSKASSY (2008),* the authors classify sentiment of firm specific news stories to predict individual accounting variables and stock returns. They use the same approach of counting positive and negative words. In a seminal work titled *When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks, Loughran and McDonald (2011)* developed an alternative financial news specific word list, along with five other word lists, that better reflected tone in financial text. They linked the word lists to 10-K filing, returns, trading volume, return volatility, fraud, material weakness, and unexpected earnings. Their dictionary has been used in many papers since and inspired many improvements in the future approaches .

However, the most important improvements in financial sentiment analysis tasks have been due to the direct adoption of research in other fields. The application of supervised learning and in particular deep learning models like LSTMs has resulted in modelling semantics better with a deeper sense of language context. This has led to better accuracy in classification and increased application.

These approaches most commonly try to model words as a vector of real numbers of pre -defined size and use these vectors to predict the tone of the entire corpus. The similarity between vectors, most commonly measured by cosine distance, serves as a representation of how close the words really are. *Word2Vec (2013) (Mikholov et. al.)* was the first such word embedding model. It revolutionized the field with the models outperforming pre - existing models in every domain.*Glove* was another such model utilising the same idea.

The latest research corresponds to use of Encoder-Decoders and transformers in the field of text classification. BERT developed by google has outperformed all the previous models. The advantage of these models is that they can be trained on domain specific literature and used further in specific corpuses.

For ex: FinBERT developed as an extension to Google's state of art BERT model for NLP was pre trained on a financial news corpus provided by Thomson Reuters. The model can then be used for further sentiment analysis tasks on a specific article(s). *The core idea behind these models is that by training language models on very large corpora and then initializing down-stream models with the weights learned from the language modeling task, a much better performance can be achieved. (Dogu Tan Araci (2019)).*

DATA

Data is gathered from online Crypto blogs specifically coindesk.com and cointelegraph.com. These 2 websites are ranked highest among 100 such websites based on Facebook fans and Twitter followers as reported by *feedspot.com*[#1]. A Scrapper was written in Python to extract the data which ranges from April 2018 to May 2021 containing 1076 unique dates from both the websites and merged into one. The price data of bitcoin was gathered from binance.com using APIs provided by them. BTCUSDT (bitcoin/tether) pair was used for the analysis.

The news articles data contains 3 fields: TITLE of the news, DATE of publication and the DESCRIPTION of the news. DESCRIPTION of the news is a kind of sub-heading containing some additional information about the title. These values where merged for unique dates containing more than 1 articles by separating

#1 : https://blog.feedspot.com/cryptocurrency_blogs/

them with a new line break (‘\n’) in python. Both TITLE and DESCRIPTION are used separately for the model. The analysis was done date-wise on the data. Proper care was taken so that the data doesn’t leak forward in time.

METHODOLOGY

The analysis step includes 4 major steps:

1. Cleaning the textual Data.
2. Transforming text to numbers.
3. Generating labels to be learnt from bitcoin price data.
4. Feeding transformed text and labels to ML models as input.
5. Measuring and reporting the performance of models.

Cleaning the textual data

It involves performing some basic operations on the data so that some dimensionality of the data is reduced and it is easier for the transformation step to work with this.

The 5 basic steps taken were:

1. Converting every word to lowercase.
2. Removing all the signs and symbols from the data (-, #, %, $, ;, , etc).
3. Removing all the digits and numbers from the text.
4. Removing more than one whitespace between words.
5. Removing all the STOPWORDS from the data[#2]

Example :

Raw Text Title

'Bitcoin Will ‘Surpass’ All-Time Price Highs by End of 2019, Says Quoine CEO\nCoinbase’s Former Fraud Lead Has Left to Join Tech Firm Twilio\nTreasury Official: Global Regulators Must Follow US Lead in Crypto Enforcement\nHyperledger Launches Cryptography Toolbox for Blockchain Developers\nSignature Bank Wins New York Approval for ‘Real-Time’ Blockchain Payments\nSwiss Regulator’s ‘Relaxed’ Fintech License Covers Blockchain Firms\nFidelity, Bitmain and More Invest $27 Million in Crypto Trading Platform ErisX\nChinese Crypto Billionaire to Help Lead Hong-Kong-Listed Blockchain

#2 The list of STOPWORDS was downloaded from popular sentiment Analysis library nltk.

Firm\nOverstock’s Medici Invests $2.5 Million in Grain Tech Firm’s Blockchain Pivot\n$35 Million: Sequoia Backs Turing Award Winner’s Blockchain Project\nTop Cryptocurrencies See Slight Gains, Bitcoin Hovers Under $4,000\nHow Traditional Financial Instruments Are Breaking Out in the World of Crypto\nCrypto Exchange OKEx Launches ‘Perpetual Swap’ Derivative Product\nJoseph Lubin: ETH Incubator ConsenSys Gets ‘Lean and Gritty’ in Competitive Market'

Cleaned Text

'bitcoin surpass time price highs end says quoine ceo coinbase former fraud lead left join tech firm twilio treasury official global regulators must follow us lead crypto enforcement hyperledger launches cryptography toolbox blockchain developers signature bank wins new york approval real time blockchain payments swiss regulator relaxed fintech license covers blockchain firms fidelity bitmain invest million crypto trading platform erisx chinese crypto billionaire help lead hong kong listed blockchain firm overstock medici invests million grain tech firm blockchain pivot million sequoia backs turing award winner blockchain project top cryptocurrencies see slight gains bitcoin hovers traditional financial instruments breaking world crypto crypto exchange okex launches perpetual swap derivative product joseph lubin eth incubator consensys gets lean gritty competitive market'

Raw Text Description

'The CEO of Japanese fintech firm and crypto exchange operator Quoine believes Bitcoin will “surpass” its all-time price highs by the end of 2019.\nOne of Coinbase\'s longest-servicing employees, risk operations manager Rees Atlas, has left the Silicon Valley cryptocurrency exchange for Twilio.\nU.S. Treasury Department Under Secretary Sigal Mandelker called for global efforts to police malicious actors\' use of cryptocurrencies.\nHyperledger has launched a new tool for blockchain developers – a modular cryptographic library aimed to reduce work duplication and bugs.\nSignature Bank is launching a blockchain-based real-time payments system in early 2019 and has just got the green light in New York.\nSwitzerland\'s Financial Market Supervisory Authority has introduced a new "relaxed" fintech license that can apply to blockchain and crypto firms.\nErisX has closed a $27.5 million Series B funding round to build a regulated crypto spot and futures market.\nVeteran Chinese crypto investor Li Xiaolai has joined a blockchain firm listed on the Hong Kong Stock Exchange as an executive director and co-CEO.\nMedici Ventures has bought a $2.5 million equity stake in GrainChain, a software firm launching its own blockchain and stablecoin.\nConflux, a scalable blockchain project with a Turing Award-winning co-founder, has raised $35 million from backers including Sequoia and Baidu.\nCrypto markets are seeing a tint of green, with just a few top coins in the red. Bitcoin is hovering under the $4,000 mark.\nWant to remove some of the volatility from your investments? Here’s a look at some of the traditional financial instruments making waves in the crypto world.\nCrypto exchange OKEx has introduced a derivative product called Perpetual Swap, allowing users to

hold positions indefinitely.\nEthereum blockchain startup and incubator ConsenSys plans to streamline and toughen its business style amid a competitive blockchain space.'

Cleaned Text

'ceo japanese fintech firm crypto exchange operator quoine believes bitcoin surpass time price highs end one coinbase longest servicing employees risk operations manager rees atlas left silicon valley cryptocurrency exchange twilio treasury department secretary sigal mandelker called global efforts police malicious actors use cryptocurrencies hyperledger launched new tool blockchain developers modular cryptographic library aimed reduce work duplication bugs signature bank launching blockchain based real time payments system early got green light new york switzerland financial market supervisory authority introduced new relaxed fintech license apply blockchain crypto firms erisx closed million series funding round build regulated crypto spot futures market veteran chinese crypto investor li xiaolai joined blockchain firm listed hong kong stock exchange executive director co ceo medici ventures bought million equity stake grainchain software firm launching blockchain stablecoin conflux scalable blockchain project turing award winning co founder raised million backers including sequoia baidu crypto markets seeing tint green top coins red bitcoin hovering mark want remove volatility investments look traditional financial instruments making waves crypto world crypto exchange okex introduced derivative product called perpetual swap allowing users hold positions indefinitely ethereum blockchain startup incubator consensys plans streamline toughen business style amid competitive blockchain space'

Transforming Text to Numbers

The most important step of any sentiment analysis task is the procedure to convert cleaned text to numeric data for the ML algorithm to understand.

4 different and popular methods were applied separately and performance was judged for each of them.

1. CountVectorizer
2. TF-IDF vectorizer
3. Word2vec
4. Glove

1. **CountVectorizer**: CountVectorizer is the simplest of tools available to accomplish the transformation. It makes a list of all the words in the corpus and then counts the frequency of each word in the particular sentence/paragraph The value of the word is either the frequency or 0. If used with constraint on the maximum number of features, it ignores the lower frequency words. In this project, a CountVectorizer with a maximum number of words as 100 was initialized. The data was separated into 3 parts train, validate and test set as 60%, 15%, 25%. The vectorizer was fitted to the training values and the test and validated set were transformed using the train set vectorizer. We obtained a matrix of num_rows * 100 as features to be fed into the ML model.

2. **TF-IDF vectorizer:** TF-IDF stands for "Term Frequency Inverse Document Frequency". It is an improvement upon the CountVectorizer and reflects how important a term is to a particular paragraph/sentence. The value of a word increases proportionally to the number of times it is there in the paragraph and is inversely proportional to the frequency of the word in the corpus. sklearn's TFIDF vectorizer is used for the purpose with a maximum number of words as 100.

3. **Word2Vec:** Word2vec is a neural network structure to generate word embedding by training the model on a supervised classification problem. The 2 methods to train a Word2Vec model are CBOW (continuous Bag of Words) and Continuous skip-gram.
   A matrix with dimensions vocabulary size * embedding size is initialized with random values and a supervised learning task is set up to predict the next word in a sentence based on a moving window. It then tries to update the respective probabilities based on the error vectors. Every word in a corpus is represented as a vector of fixed size. The vector elements try to capture some dependence structure. We can either use a pre-trained word2vec model or train our own model using the training data. Since our input data is very limited it makes sense to use a pre-trained model.
   In this project a pre-trained model called "google-news-300" is used. It stores the words as a vector of 300 features and has been pre-trained on 100 billion words from part of google news dataset.

4. **Glove :** Another pre-trained model called “glove-twitter-25” which was trained on 2 billion tweets with 27 billion tokens and 1.2 million vocab size data from twitter has been used. It stores each word as a vector of size 25. The performance of both is compared.

Generating labels

The direction of 1 day future return of the bitcoin price is used as the label.If the price went up next day the label is 1 else it is 0.There is a slight imbalance in the dataset with 53% as positive and 47% as negative.

Modelling

The whole data is divided randomly into 3 parts: train, validation and test set as (60% ,15%,25%). The embeddings from step 2 are used as the features with labels as mentioned above.

We have 4 embeddings (1 countVectorizer, 1 tf-idf vectorizer and 2 word2vec models). Four different models have been used for the classification task: LogisticRegression, SVClassifier, RandomForestClassifier and MLP Classifier. This results in 16 different combinations. The model fitting and hyperparameters optimization has been done using only the train and validation set. The test set was only used for fitting the best model and reporting the results.

Hyperparameter optimization was done with gridsearchCV with cv set of 5. Along with accuracy and confusion matrix, returns of a hypothetical strategy with predictions as signals is reported. Sharpe ratio of the strategy is also reported after accounting for transaction costs of 25bps per day[#1].

RESULTS

BaseLine Model

A logistic regression classifier with Countvectorizer without any

#1 This has been done after taking into account binance transaction costs(10 bps per trade and 5bps slippage).

hyperparameter tuning was chosen as a baseline model. Logistic Regression model from sklearn package was used with default parameters and no cross validation.

The performance of the baseline model was good achieving an accuracy of 67% on the training set and 57% on the validation set.

TRAIN SET

LABEL	precision	recall	f1-score	support
0	0.67	0.58	0.62	297
1	0.62	0.59	0.61	348
accuracy			0.67	645
macro avg	0.67	0.67	0.67	645
weighted avg	0.67	0.67	0.67	645

VALIDATION SET

LABEL	precision	recall	f1-score	support
0	0.50	0.53	0.51	70
1	0.62	0.59	0.61	91
accuracy			0.57	161
macro avg	0.56	0.56	0.56	161
weighted avg	0.57	0.57	0.57	161

Next, we train a TF-IDF model with logistic regression as the ML model. This also had an accuracy similar to the CountVectorizer with train accuracy 67% and validation accuracy 60%.

The Next table displays the training set accuracy of the 4 models applied on the four feature generating schemes on the ‘TITLE’ of the news.We achieve strong

in-sample performance for neural networks. RandomForestClassifier comes in second.

TRAIN ACCURACY	CountVectorizer	Tf-IDF Vectorizer	Google-news-300	Glove-twitter-25
LogisticRegression	**0.67**	**0.59**	**0.57**	**0.46**
SVC Classifier	**0.51**	**0.44**	**0.46**	**0.44**
RandomForestClassifier	**0.56**	**0.58**	**0.64**	**0.59**
MLP Classifier	**0.86**	**0.85**	**0.97**	**0.55**

VALIDATION ACCURACY	CountVectorizer	Tf-IDF Vectorizer	Google-news-300	Glove-twitter-25
LogisticRegression	**0.57**	**0.55**	**0.51**	**0.43**
SVC Classifier	**0.49**	**0.53**	**0.43**	**0.53**
RandomForestClassifier	**0.55**	**0.52**	**0.48**	**0.5**
MLP Classifier	**0.54**	**0.58**	**0.49**	**0.57**

TEST ACCURACY	CountVectorizer	Tf-IDF Vectorizer	Google-news-300	Glove-twitter-25
LogisticRegression	0.49	0.48	0.43	0.5
SVC Classifier	0.46	0.51	0.5	0.51
RandomForestClassifier	0.56	0.56	0.44	0.47
MLP Classifier	0.46	0.51	0.49	0.51

The performance of the models fizzes out in the validation and test test. Only RandomForests perform better than random 2 times on CountVectorizer and TF-IDF embedding with accuracy of 56%. Rest of the models are worse .

The models were also applied on the ‘DESCRIPTION’ of the data. The results were as follows :

TRAIN ACCURACY	CountVectorizer	Tf-IDF Vectorizer	Google-news-300	Glove-twitter-25
LogisticRegression	0.66	0.61	0.52	0.46
SVC Classifier	0.5	0.73	0.46	0.54
RandomForestClassifier	0.56	0.66	0.69	0.69
MLP Classifier	0.88	0.54	0.97	0.56

VALIDATION ACCURACY	CountVectorizer	Tf-IDF Vectorizer	Google-news-300	Glove-twitter-25
LogisticRegression	0.54	0.52	0.51	0.43
SVC Classifier	0.5	0.5	0.43	0.57
RandomForestClassifier	0.52	0.55	0.45	0.51
MLP Classifier	0.53	0.52	0.5	0.61

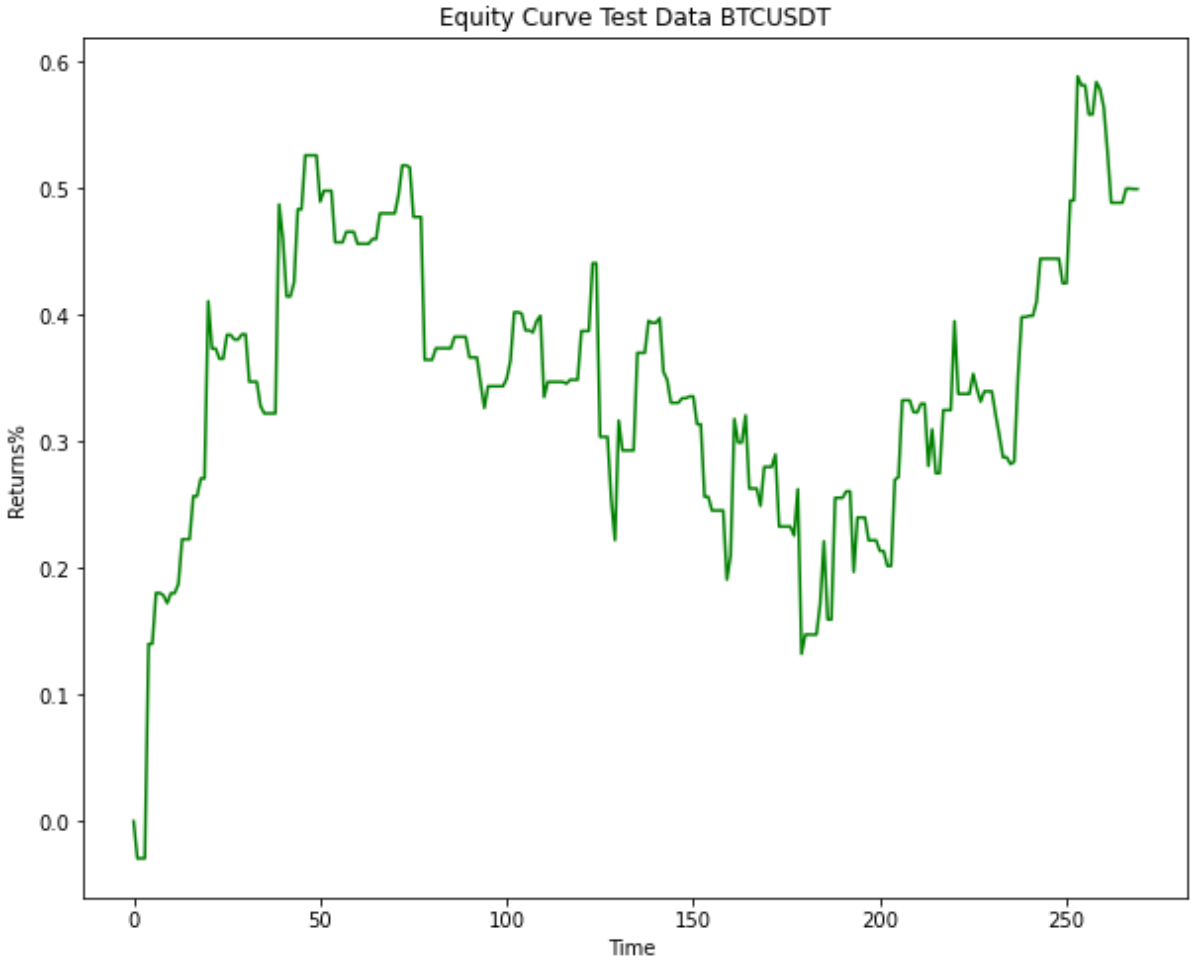
TEST ACCURACY	CountVectorizer	Tf-IDF Vectorizer	Google-news-300	Glove-twitter-25
LogisticRegression	0.51	0.46	0.42	0.5
SVC Classifier	0.5	0.5	0.51	0.5
RandomForestClassifier	0.49	0.5	0.45	0.5
MLP Classifier	0.5	0.52	0.51	0.53

Here, again MLPClassifier performed amazingly in train datasets but failed to perform in validation and test sets even after extensive hyperparameter optimization.The performance of the models is worse than that of 'TITLE' category based on accuracy metric with accuracy nearly always less than 50%. Hence we can safely say that the models failed to learn anything of value from the data. The comparison of embeddings paint another interesting picture. Roughly looking, we can say that the two word2vec embeddings failed to outperform their simpler counterparts in both datasets.The glove – twitter dataset faired slightly better than google -news -300.Within the two counting

based embeddings, TF-IDF vectorizer performed better than CountVectorizer in both datasets.

Our best models are RandomForestClassifier with CountVectorizer and TF-IDF embedding with accuracy of 56% each on the title dataset. Here are the pnl and sharpe ratio of each of these classifiers.

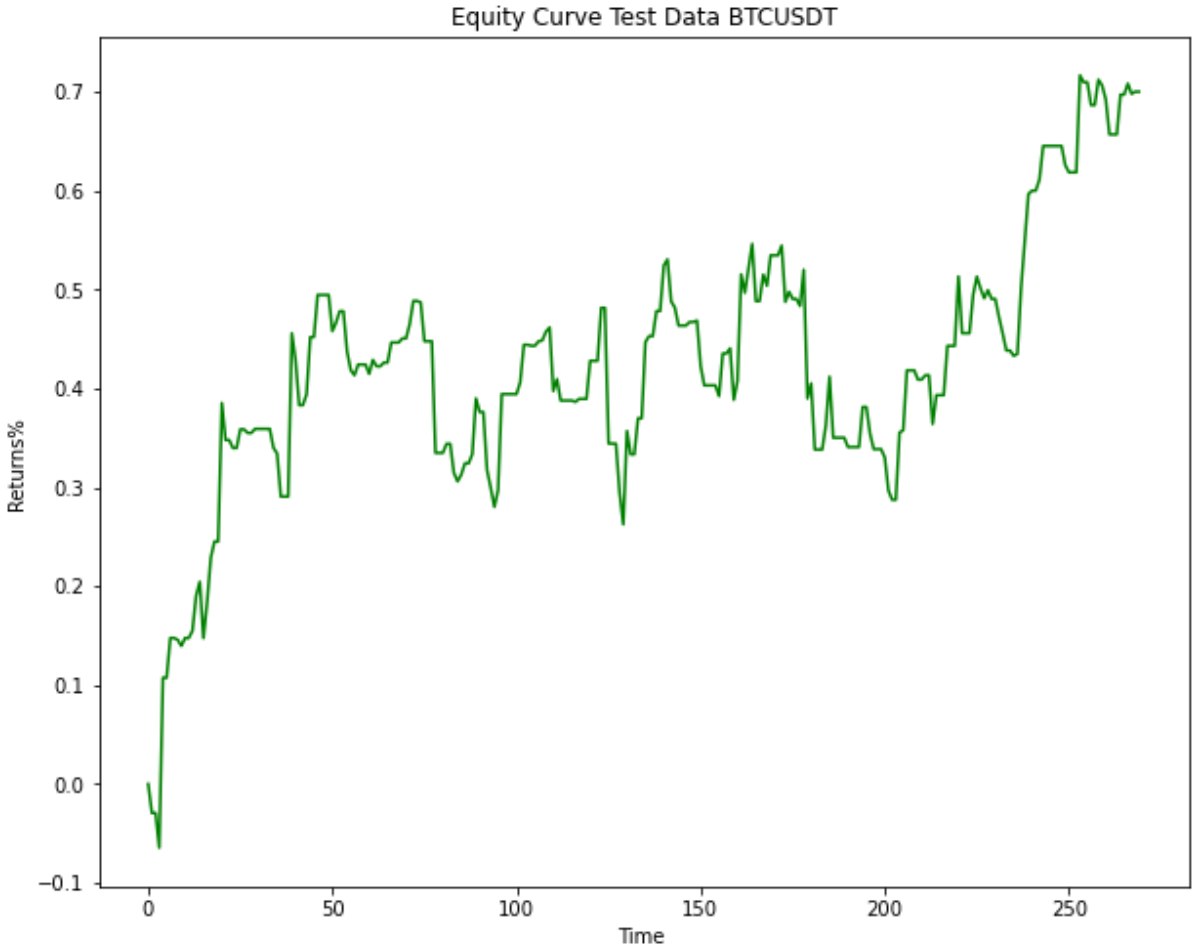
CountVectorizer



ANNUALIZED RETURNS : 47%

ANNUALIZED sharpe : 0.90

TF-IDF Vectorizer



ANNUALIZED RETURNS : 65.36%

ANNUALIZED sharpe : 1.22

The return and risk profile for Tf-IDF embedding looks much better than that of CountVectorizer. Both models have good returns and attractive sharpe ratios.

CONCLUSION

The performance of the models on this dataset was below average. All the models barring two performed poorly in the test and validation set. The performance of the neural network classifier was amazing in the train set but even after extensive hyperparameter optimization the model failed to generalize out of sample. The poor performance may be attributed to a couple of things.

One reason could be the limited data set consisting of only around thousand instances. Another limitation of the above exercise is the use of only 1 length word for tokenization, including bigrams and n-grams may result in better model performance. A second major problem may be due to the domain itself, application of ML techniques to financial markets hasn't been as successful as some other fields and the accuracy of complex models don't reach 60% most of the time. It is also possible that the data on a standalone basis has no predictive power for bitcoin. This would warrant an application of the model in conjunction with other types of data like price and volume and other indicators. Limited availability of data can be mitigated partially by using models which have been pre-trained on financial corpus. FinBERT as mentioned earlier, trained entirely on financial literature can be used for the task. Future work in this field would be the application of FinBERT and the latest Transformers based model to various types of financial data including corporate updates and announcements and stock message boards data.

REFERENCES

1. *FinBERT: Financial Sentiment Analysis with Pre-trained Language Models by Dogu Tan Araci.*
2. *When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks by TIM LOUGHRAN and BILL MCDONALD*
3. *Giving Content to Investor Sentiment: The Role of Media in the Stock Market Paul C. Tetlock*
4. *Cryptocurrency Price Prediction Using News and Social Media Sentiment by Connor Lamon, Eric Nielsen, Eric Redondo.*
5. *Price Movement Prediction of Cryptocurrencies Using Sentiment Analysis and Machine Learning by Franco Valencia, Alfonso Gómez-Espinosa and Benjamín Valdés-Aguirre*
6. *Deep Learning Based Text Classification: A Comprehensive Review by Shervin Minaee, Snapchat Inc Nal Kalchbrenner, Google Brain, Amsterdam, Erik Cambria, Nanyang Technological University, Singapore Narjes Nikzad, University of Tabriz Meysam Chenaghlu, University of Tabriz Jianfeng Gao, Microsoft Research, Redmond*
7. *Efficient Estimation of Word Representations in Vector Space by Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean*