MICROBLOG GUIDED CRYPTOCURRENCY TRADING AND FRAMING ANALYSIS

By

Anna Paula Pawlicka Maule

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Computer Science – Master of Science

2020

ABSTRACT

MICROBLOG GUIDED CRYPTOCURRENCY TRADING AND FRAMING ANALYSIS

By

Anna Paula Pawlicka Maule

With 56 million people actively trading and investing in cryptocurrency online and globally, there is an increasing need for an automatic social media analysis tool to help understand trading discourse and behavior. Previous works have shown the usefulness of modeling microblog discourse for the prediction of trading stocks and their price fluctuations, as well as content framing. In this work, I present a natural language modeling pipeline which leverages language and social network behaviors for the prediction of cryptocurrency day trading actions and their associated framing patterns. Specifically, I present two modeling approaches. The first determines if the tweets of a 24-hour period can be used to guide day trading behavior, specifically if a cryptocurrency investor should buy, sell, or hold their cryptocurrencies in order to make a trading profit. The second is an unsupervised deep clustering approach to automatically detect framing patterns. My contributions include the modeling pipeline for this novel task, a new dataset of cryptocurrency related tweets from influential accounts, and a transaction volume dataset. The experiments executed show that this weakly-supervised trading pipeline achieves an 88.78% accuracy for day trading behavior predictions and reveals framing fluctuations prior to and during the COVID-19 pandemic that could be used to guide investment actions.

This thesis is dedicated to those that seek financial freedom.

ACKNOWLEDGEMENTS

First, I would like to thank my advisor, Dr. Kristen Johnson, for believing in me and allowing me to learn from her indispensable research expertise. Her feedback and support allowed me to thrive, stay motivated, and push my knowledge boundaries further.

I would like to thank my mother, Professor Agnieszka Pawlicka, my uncle, Dr. Jakub Pawlicki, my grandparent, Professor Grzegorz Pawlicki, and my grandmother M.D. Halina Pawlicka for being role models and always emphasizing the importance of continuous learning. I would also like to thank my dad, Cassio Maule, that despite the lack of any academic titles, taught me flawlessly calculus and linear algebra over lunch, and also taught me how to be a caring humble human being.

I would like to thank my boyfriend, Kasper Standio, for sharing his passion for cryptocurrencies with me and spending countless hours discussing ideas and sharing his knowledge on this topic. He also gave me some insightful feedback on my research and helped annotate data for this project.

I would like to thank my lab mate, Zachary Yarost, for annotating the dataset for this work. I would also like to thank my co-workers from TechSmith for showing their support during my Master's program. My co-workers went above and beyond to accommodate my work schedule changes every semester.

Finally, I would like to thank my committee members, Dr. Parisa Kordjamshidi, and Dr. Jiayu Zhou.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

NN Neural Network

RNN Recurrent Neural Network

LSTM Long Short-term Memory

BERT Bidirectional Encoder Representations from Transformers

LDA Latent Dirichlet Allocation

BTC Bitcoin

ETH Ethereum

WHO World Health Organization

COVID-19 Coronavirus Disease 2019

NASDAQ National Association of Securities Dealers Automated Quotations

CHAPTER 1

INTRODUCTION

1.1 Motivation

Beginning with the 2008 introduction of Bitcoin (BTC), a cryptocurrency for a Peer-to-Peer cash system, the use of cryptocurrencies and their corresponding blockchains have become increasingly popular. In 2019, the number of Americans owning cryptocurrency doubled from 7% in 2018 to 14%, representing about 35 million people trading and investing with cryptocurrency [55].

This increase is largely due to the capability of cryptocurrency to improve various applications ranging from increased security of smart contracts to facilitating less expensive and faster cross-border international payments. Another contributing factor to this growth is that digital coins fulfill the property of storing value similar to any other *fiat currency*, which is a government-issued currency that is not backed by physical commodities, e.g., the American dollar or euro. Finally, cryptocurrency popularity has been boosted due to its high day trading volume. In October 2020, the combined worth of the top 10 cryptocurrencies was $340 billion, with Bitcoin accounting for $250 billion of this amount. Since January 2020, the median day trading volume of Bitcoin has been $30 billion. To put this in perspective, the trading volume of Alphabet Inc. (the parent company of Google) in the past 3 months has been $2.75 billion, while Amazon Inc. has a trading volume average of $15.6 billion per day – over 6 times more than that of Google, but still around 0.5 times less than the BTC daily volume. [1]

Cryptocurrencies were born on the internet, gained their visibility through online and social media coverage, and many investors follow the advice of well-known cryptocurrency experts on Twitter to guide their personal investment strategies [51]. Because cryptocurrency prices can fluctuate quickly, resulting in real-life financial gains or losses, models that can rapidly analyze trending discourse on Twitter can be harnessed to guide and benefit investors.

[1] https://finance.yahoo.com/quote/GOOGL/; https://finance.yahoo.com/quote/AMZN

Additionally, work in computational linguistics and social sciences have shown the benefit of studying *framing*, i.e., how someone spins a topic to sway the opinion of the public. Framing in Twitter discourse can be used to understand social phenomena, such as political maneuvering or epidemiology coverage. However, little work exists studying the relationship between economic framing and stock or cryptocurrency trading, especially during times of economic stress.

Currently, it is estimated that the COVID-19 pandemic will negatively impact the global economy by hindering economic growth worldwide between 3.0% to 6.0% and potentially causing global trade to fall up to 32% [21]. Similar to the pandemic's effect on Wall Street (i.e., the New York Stock Exchange and NASDAQ), the cryptocurrency market reflected a drastic 47.8% drop on March 12, 2020. This drop occurred one day after the World Health Organization (WHO) announced that COVID-19 could be characterized as a pandemic. This same pattern followed stocks worldwide within a similar time frame. This trend led to the hypothesis that how people frame day trading behaviors (e.g., buy or sell) would be a useful predictive feature in understanding cryptocurrency trading.

1.2 Contributions

To this end, I have developed a dual cryptocurrency day trading behavior modeling pipeline that leverages language and social network behavior extracted from tweets to: (1) implement a weakly-supervised predictive model that predicts investment action, specifically, whether to buy, sell, or hold cryptocurrency based off of discussions from tweets within a 24-hour period, and (2) implement an unsupervised deep-learning clustering model to determine the underlying framing patterns used to discuss these cryptocurrency investment actions. Additionally, my contributions include a cryptocurrency-related tweets dataset and Bitcoin historical transaction volume dataset.[2] My models show a distinction between how day trading is framed before and during the pandemic as well as a strong correlation between these different frames and the buying or selling of cryptocurrency.

[2]Datasets and code will be made publicly available after conference publication.

1.3 Overview

This thesis is organized into seven chapters. Chapter 1 has introduced the motivation of this work and the contributions. Chapter 2 gives a brief overview of machine learning, natural language processing, and cryptocurrency concepts that were utilized for the development of this work. Next, Chapter 3 describes related works such as online discourse analysis, stock market prediction, optimal historical data collection, cryptocurrency price prediction, and framing theory in microblogs. Chapter 3 concludes with a section comparing the novel contributions of this thesis to the related works. The Data Annotation chapter (Chapter 4) focuses on describing the cryptocurrency tweets and prices collection, as well as what pre-processing steps were applied to this newly generated tweets dataset. Chapter 5 explains the models and feature engineering utilized for both day trading behavior prediction and discourse framing clustering. In the subsequent Chapter 6, the experimental setup, trials, and accuracy of the developed models are presented. Chapter 7 analyzes the qualitative results of the frames before and during the pandemic, and also inspects the correlation between the frames and momentum patterns. This thesis is finalized with a conclusion and future work discussion in Chapter 8.

CHAPTER 2

BACKGROUND INFORMATION

This chapter provides an overview of the background knowledge used in the development of this thesis. The core areas incorporated into this work are machine learning, natural language processing, and cryptocurrencies, with this chapter divided into those three sections respectively. While each of these areas covers a vast amount of knowledge, each section of this chapter focuses on clarifying only the concepts from each area utilized for the development of this thesis.

2.1 Machine Learning Algorithms

Machine learning algorithms are able to learn from a dataset. Mitchell defines machine learning as the following: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" [50]. Machine learning algorithms and models can be supervised, unsupervised, or weakly-supervised. Supervised learning is when the machine learning algorithm is provided with a fixed set of features and known labels for the input and output, and the algorithm learns a mapping from input features to output prediction. Conversely, unsupervised learning approaches do not require prior knowledge about the features or labels of the dataset, instead deducing this information on their own. In this section, supervised algorithms, such as Naives Bayes, and unsupervised models (e.g., Autoenconders and Deep Clustering) are explained.

2.1.1 Naive Bayes Classifier

The Naives Bayes classifier is a probabilistic classifier based on the Bayes' theorem [41]:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)} \tag{2.1}$$

In this theorem A and B are events, and $P(B) \neq 0$.

Under this theorem the following hold:

- $P(A \mid B)$ is a conditional probability. This probability can be read as the following: What is the likelihood of A happening given B happened?
- $P(B \mid A)$ is also a conditional probability. It can be read as the following: What is the likelihood of B happening given A happened?
- $P(B)$ is the probability of observing event B. This is a marginal probability.
- $P(A)$ is the probability of observing event A. This is also a marginal probability.

For the classification framework, Bayes's rule can be written in the following form:

$$P(c_k \mid x) = P(c_k) \times \frac{P(x \mid c_k)}{P(x)} \tag{2.2}$$

Where c_k are the documents (classes) and x is a set of features, such as words. The formula can further be simplified by ignoring the probability of $P(x)$, given they will be the same when computing the probability of $P(c_k \mid x)$ for every c_k.

$$P(c_k \mid x) = P(c_k)P(x \mid c_k) \tag{2.3}$$

Estimating $P(x \mid c_k)$ can be complex because there are a vast possibility of values for $x = (x_1, x_2, ..., x_i)$. Therefore, it can be assumed that the distribution of x conditional on c_k can be expressed in the following manner for all values of c_k [47]:

$$P(x \mid c_k) = \prod_{j=1}^{d} P(x_j \mid c_k) \tag{2.4}$$

Then the equation can be expressed as follows:

$$P(c_k \mid x) = P(c_k) \prod_{j=1}^{d} P(x_j \mid c_k) \tag{2.5}$$

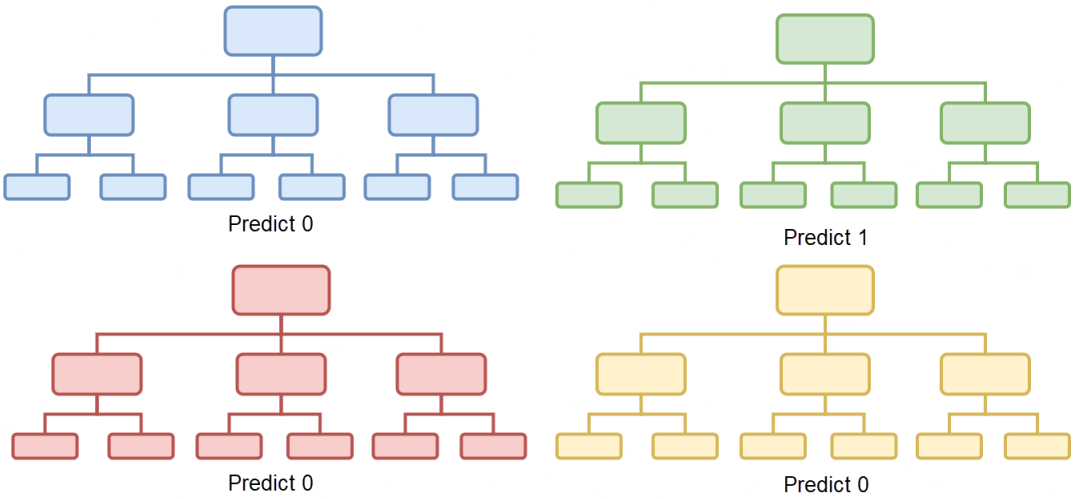


Figure 2.1: Random Forest Classifier. A random forest composed of these four decision trees would have a final prediction of Class 0.

The formula above is used to compute the probability of c_k given X for all C. The class c_k with the highest probability, $P(c_k \mid x)$, is the class selected by the classification model. The final formula used in the Naive Bayes model in this work is listed below.

$$C = argmax_C P(C) \prod_{j=1}^{d} P(x_j \mid C) \tag{2.6}$$

2.1.2 Random Forest Classifier

The Random forest classifier is an ensemble of several decision trees. A decision tree is a machine learning algorithm, used for regression and classification, where the nodes represent the features (classes) and the leaf nodes (the last node of a tree branch) is the output of the model. A random forest model aggregates the effort of several deep decision trees and then averages their result (similar to k-fold cross validation) with the goal of reducing the variance and keeping the bias low. Compared to random forest, regular decision trees have low bias, but high variance [33].

Figure 2.1 presents an example of the random forest model making a prediction. There are four decision trees, where three out of the four have predicted class zero, and only one tree predicted class one. Since the majority of the decision trees predicted class zero over class one, the random

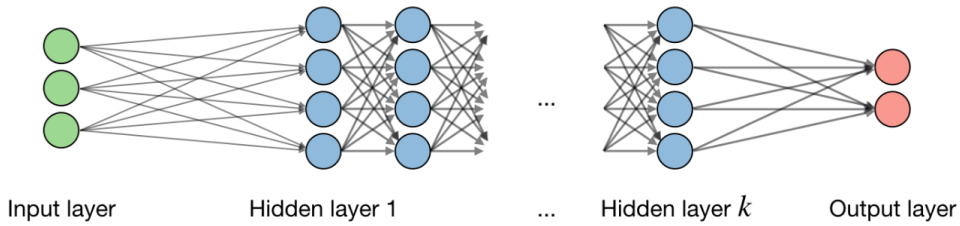


Figure 2.2: Neural Network Architecture [9].

forest classifier predicts class zero.

An important requirement of random forest is that the decision trees within the random forest must be uncorrelated. Random forest guarantees that the decision trees are uncorrelated by using the bagging (bootstrap aggregation) technique. Bagging consists of randomly selecting samples without excluding those samples for the next tree composition.

2.1.3 Neural Network

Dr. Robert Hecht-Nielsen, the pioneer in artificial neural networks (ANN), defines ANN as ". . . a computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs." In other words, neural networks are non-linear statistical models that are built of layers (input layer, a number of hidden layers, output layer) with the objective to find patterns in complex datasets. The following text briefly describes each component of the neural network architecture.

Architecture. The neural network architecture has an input layer, hidden layers, and an output layer. Each neuron (node) is connected to all the nodes of the next layer as shown in Figure 2.2.

Neuron. The artificial neuron is inspired by the neuron of a human brain as shown in Figure 2.3. Similar to the way that the neural neuron is the basic unit in the nerve system, the artificial neuron is the smallest unit in the computational artificial network. The artificial neuron has incoming inputs with their respective weights (w_0x_0, w_1x_1, ..., w_ix_i), a cell body that will sum all the inputs

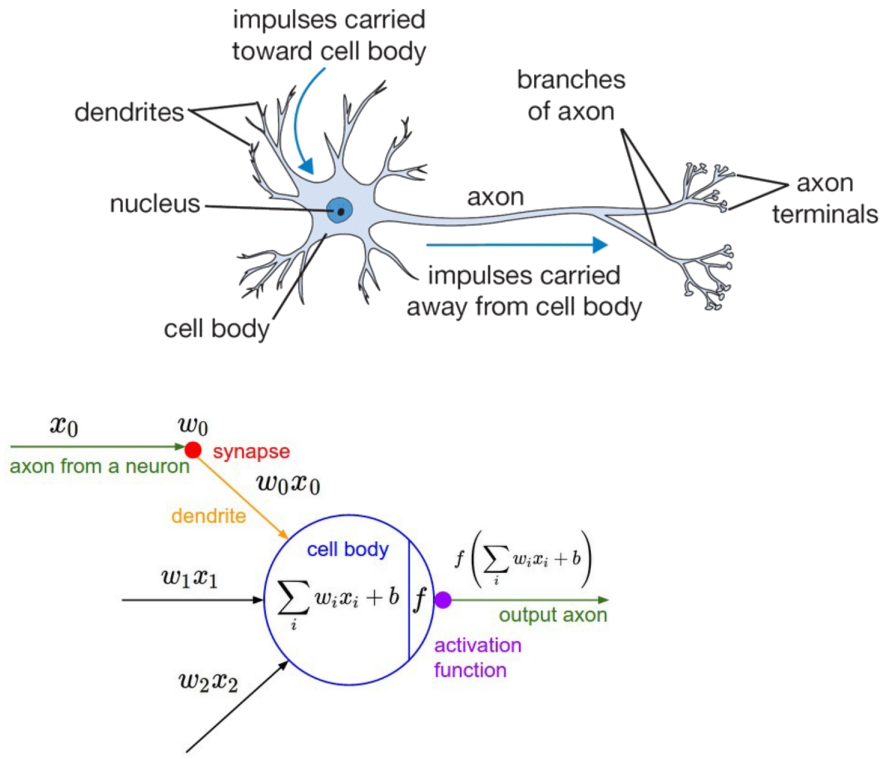


Figure 2.3: Biological Neuron (top) and Artificial Neuron (bottom) [27].

together and calculate, by using an activation function, if the impulse should be fired through the node's output axon.

Activation Function. The activation of the artificial neuron is the abstraction of firing a stimulus on a biological neuron. In a neural network these are efficient mathematical functions, such as the sigmoid or tanh functions shown in Figure 2.4, which can determine if the neuron should "fire" or not. For example, depending on the function, it can return 0 to indicate the neuron should not activate and 1 to represent it should activate. Activation represents passing the current value on to the next layer of the neural network.

Gates. Gates are a way to optimally let information through a cell state. In order to achieve that they are composed of a sigmoid neural network layer and a point wise multiplication operation [5].

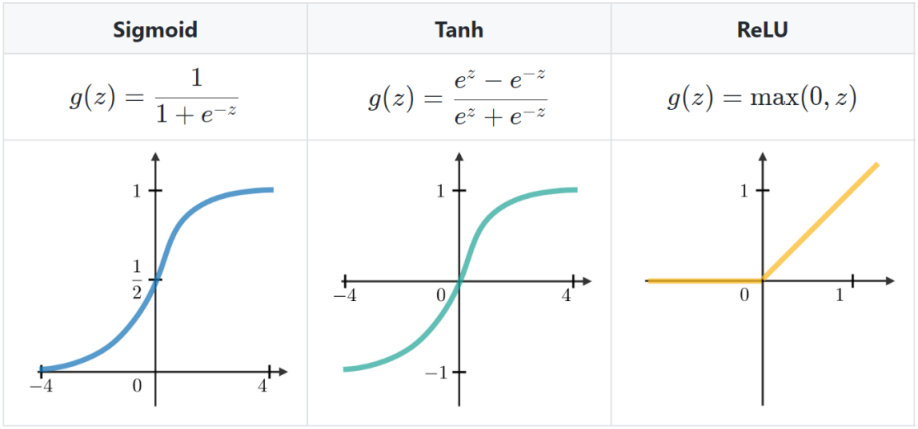


Figure 2.4: Common Neural Network Activation Functions [9].

Learning Rate. The learing rate is a hyperparameter that controls the percentage of change of the model in reaction to the estimated error every time the model weights are updated [15]. Choosing a learning rate that is too low may result in a slow training process. However, picking a high learning rate can result in the model not being able to train to find the optimal weight values for the input resulting in poor performance.

Cross-Entropy. Neural networks typically have plateaus in the learning rate, meaning that the loss point gets stuck in a local minima (Figure 2.5). Cross-Entropy is used to address this learning slowdown by replacing the quadratic cost with the cost function below [40].

$$-\Sigma_{c=1}^{N} y_{o,c} \log(1-p) \tag{2.7}$$

Entropy is the quantity of bits that is necessary to transmit a randomly selected event from a probability event. A shifted (skewed) distribution contains a low entropy, while a distribution that has equal distribution across its event has a large entropy [14]. In machine learning, cross-entropy and log loss are the same when calculating the error rates between 0 and 1. A perfect model would have a log loss of 0, and predicted probability of 100%. The log loss equation (Equation 2.8) takes in y, a binary representation (0 or 1), to indicate if the class label c is the correct classification of observation o. The input variable p represents the predicted probability observation o is of the

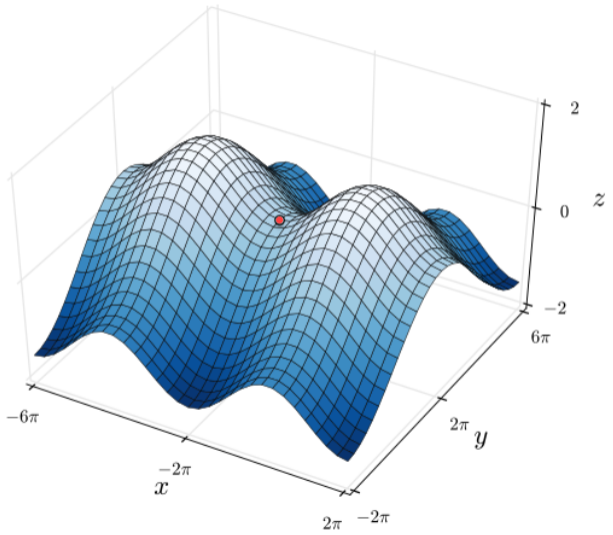


Figure 2.5: Local Minima Example. Visualization of a loss point in between two local maxima [20].

predicted class c [4].

For a binary classification, where $N = 2$, the log loss function can be represented as:

$$L(y, p) = -[y \log(p) + (1 - y) \log(1 - p)]] \tag{2.8}$$

Backpropagation. Backpropagation is a mathematical procedure that allows a neural network model to efficiently evaluate the gradient of the error function used in the neural network. The gradient information can speed up the rate at which the minima of the error function is found [11], consequently resulting in a neural network with optimal weights that minimize loss.

Weights. Figure 2.3 illustrates how the artificial neuron has incoming and outgoing connections with a weight w_i assigned to each. These weights represent the relevance of a particular connection, i.e., the importance of that input or output in the model [11]. Algorithm 1 details the pseudo-code for the weight update procedure.

Algorithm 1: Neural Network Weights Update Procedure

$neuralNetwork \leftarrow InitializeNeuralNetwork()$
$batch \leftarrow SelectBatch(trainingData)$
$loss \leftarrow neuralNetwork.ForwardPropagation(batch)$
$gradient \leftarrow neuralNetwork.Backpropagate(loss)$
$neuralNetwork.UpdateWeights(gradient)$

Dropout. Dropout is a neural network technique that aims to prevent overfitting of the training data by dropping out neurons and its connections, similar to pruning, in the neural network during training [34].

2.1.3.1 Recurrent Neural Network

Recurrent Neural Network (RNN) is a type of neural network where the connections between neurons form a directed graph along the temporal sequence of the input. This directed graph structure allows the connections to propagate the information forward and backwards. In other words, it is a network that has some cyclic connections between neurons [40]. With these properties the RNN exhibits temporal dynamic behavior. Since RNNs are derived from simple feedforward neural networks, they are able to use their memory (internal state) to process several length sequences of inputs [53]. This architecture is very powerful, and has been proven to be useful when applied to spoken and written language problems, however it is more challenging to train [40].

2.1.3.2 Long Short-Term Memory

Long Short-Term Memory (LSTM) is a type of RNN that adds "forget gates" to prevent the vanishing gradient problem. This model also prevents backpropagated errors from disappearing or exploding. The main characteristic of an LSTM is the ability to learn tasks that require memories of events that occurred at least thousands of discrete time steps prior [35].

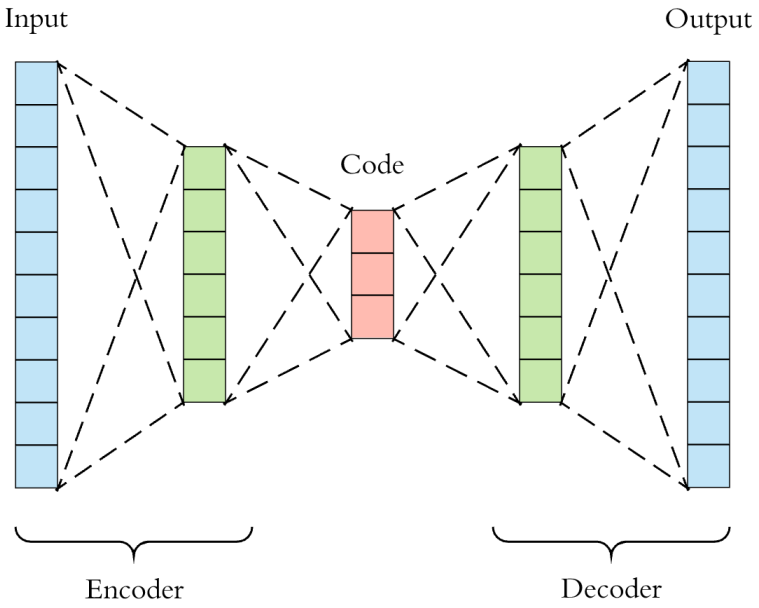


Figure 2.6: Autoencoder Structure [23].

2.1.4 Autoencoder

Autoencoder is a type of artificial neural network that aims to learn the essence, or representation (encoding), of a dataset by identifying and removing the noise signals in an unsupervised manner. In other words, the autoencoder is a model that learns how to reduce the dimensionality of a dataset, as well as how to rebuild the original dataset from a compressed state [29].

2.1.5 K-means Clustering

K-means is a clustering algorithm that given n number of clusters, the algorithm randomly selects the initial n centroids and through a predetermined maximum number of iterations, or until it reaches convergence, assigns the data points to the closest centroid. After each iteration the cluster's centroid gets recalculated. The convergence is achieved when no data points change their

cluster assignment compared to the prior iteration [30].

Algorithm 2: K-means Pseudo-Algorithm

```
Result: collection of clusters with their respective data points
centroidsList = randomlySelectCentroids(n, dataPoints);
while i < maxIteration do
    swapCount = 0;
    forall dataPoints do
        closestCluster = minDistance(dataPoint, centroidsList);
        if dataPoint.cluster != closestCluster then
            dataPoint.cluster = closestCluster;
            swapCount++;
        end
    end
    if swapCount == 0 then
        break;
    end
    centroidsList = RecalculateCentroids(dataPoints);
end
```

Formal Definition. Given a set of points $(x_1, x_2, x_3, ...x_n)$ where each point is a d-dimensional real vector. The k-means algorithm objective is to partition the data into k cluster sets $C = \{C_1, C_2, C_3, .., C_k\}$, where $k \leq n$, to minimize the variance (sum of squares) within clusters [54].

The ultimate objective is minimize is:

$$arg_c\, min \sum_{i=1}^{k} \sum_{x \epsilon C_i} \parallel x - \mu_i \parallel^2 = arg_c\, min \sum_{i=1}^{k} \mid C_i \mid Var\, C_i \tag{2.9}$$

where μ_i is the mean of the points, the centroid, in C_i. This objective function can further be simplified to:

$$J = \sum_{i=1}^{k} \sum_{j=1}^{n} \parallel x_i^{(j)} - c_j \parallel^2 \tag{2.10}$$

where k is the number of clusters C and n is the number of observations, or points, x.

2.1.6 Deep Embedded Clustering

Deep Embedded Clustering (DEC) is an unsupervised deep learning clustering algorithm that learns a mapping from the data space X to a lower-dimensional feature space Z, where it will iteratively optimize the clustering with the help of parameter initialization using an autoenconder [71].

There are two main steps to the DEC approach. The first step is parameter initialization using a deep autoencoder and the second step is parameter optimization (clustering) [71]. Instead of clustering directly on the original data space, the autoenconder transforms the raw data with a nonlinear mapping. Furthermore, the new latent feature space, generated by the autoenconder, is usually considerably smaller than the original space. After the input data is compressed by the autoenconder, the DEC layer is initialized with the centroids of k-means obtained from the new feature space Z [28] for the cluster assignment process. This DEC self training step is achieved by having a distribution that strengthens the prediction, by emphasizing data points that have higher confidence and blocking large clusters from altering the hidden feature space. In order to learn from the high confidence assignments, several iterations of the target distribution are necessary. After a maximum threshold of iterations occurs the clustering model will minimize the Kullback-Leibler divergence[1] loss between the target distribution and the clustering output.

[1]The Kullback-Leibler divergence is a measure of how one probability distribution is different from a second distribution. It is also known as relative entropy.

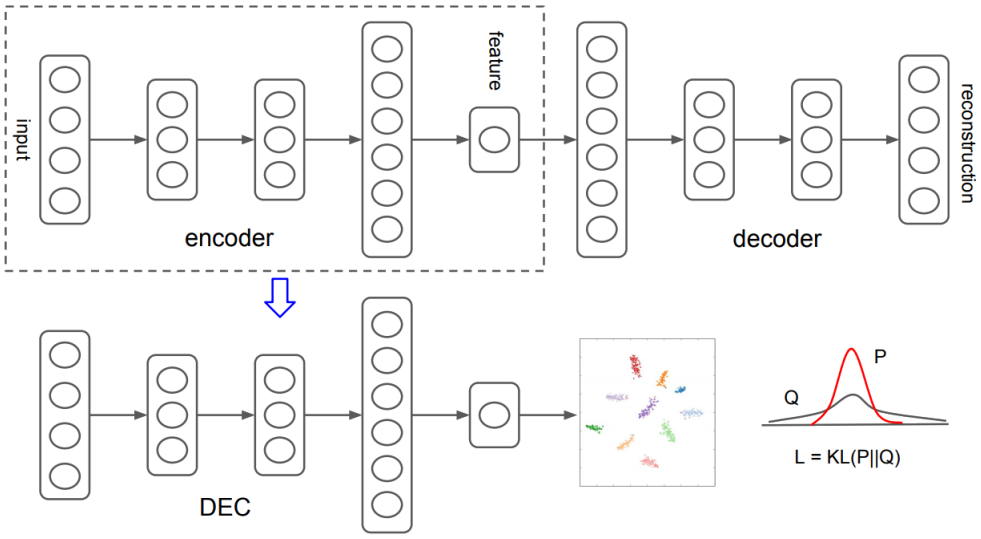


Figure 2.7: DEC Structure [71].

2.1.7 Conditional Random Field

Conditional Random Field (CRF) is a discriminative graphical model that implements dependencies between predictions. These models are used for pattern recognition or tasks where the contextual information of the neighbors impact the current prediction [46]. CRFs are best suited for sequential prediction tasks such as gene sequencing or image segmentation processing. In natural language processing tasks, CRFs are useful for Part of Speech (POS) tagging[2] and named entity recognition (NER)[3]. The most commonly used graph for NLP tasks is a linear chain, which is known for implementing sequential dependencies in the predictions [3].

2.1.8 XGBoost

XGBoost stands for "Extreme Gradient Boosting" and is a distributed gradient boosting tree library engineered to be highly efficient, portable, and flexible. It implements machine learning algorithms

[2]POS tagging is a process to tag words of a sentence based on their part of speech (e.g., verb, noun, adjective, proper noun) and their context.

[3]NER is a task in NLP that seeks to find and classify named entities such as organization names, locations, proper names, etc.

under the Gradient Boosting framework. XGBoost provides a parallel tree boosting that solves several data science applications in an efficient and accurate way [6].

2.2 Natural Language Processing

Natural Language Processing (NLP) is a junction of computer science, linguistics, and artificial intelligence that focuses on understanding various aspects of human language and its interactions through the help of computer automation and machine learning modeling. Some examples of natural language processing applications are machine translation, spelling and grammar correction, extracting meaning from text, and many others. The focus of this section is to clarify exclusively the natural language processing concepts that are applied in this thesis.

2.2.1 N-gram Representations

N-gram is a continuous sequence of N samples of text [40]. These samples can be syllables, words, phonemes, or letters. For instance if the text sample is words, then the sentence "This thesis studies cryptocurrency framing on Twitter" is a 7-gram, while "I trade cryptocurrency" is a 3-gram (trigram). The most commonly studied N-gram representations in NLP tasks are the unigram, bigram, and trigram.

2.2.2 Bag of Words

Bag of Words (BoW) is a basic representation of the words which occur in a document or dataset. In order to implement Bag of Words the following are needed: a dictionary of accepted words, a measurement of frequency, and the assumption that the positions of the words are irrelevant [40]. This representation is often used as a simple baseline for NLP model comparisons.

2.2.3 Bidirectional Encoder Representations from Transformers

Bidirectional Encoder Representations from Transformers (BERT) is a recently developed language representation model. BERT is designed to pre-train deep bidirectional representations from

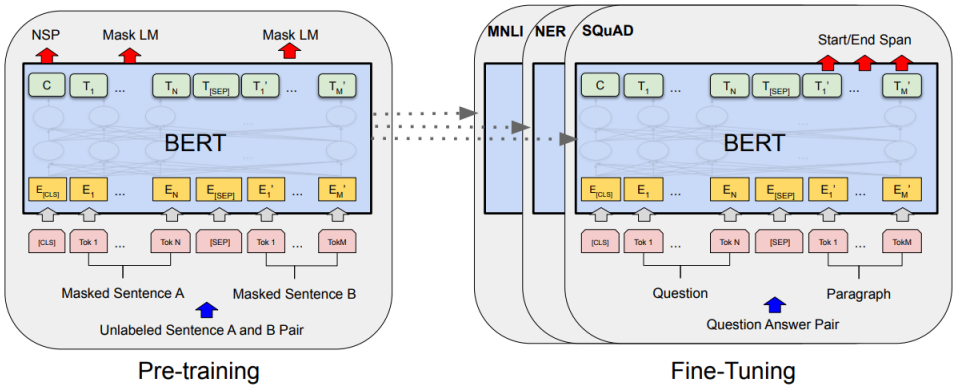


Figure 2.8: Comparison of a Pre-trained BERT model and Fine-tuned BERT model [24].

unlabeled text by simultaneously conditioning on both the left and right context in all layers. Therefore, the pre-trained BERT model can be easily fine-tuned (Figure 2.8) with just one additional output layer to create models for a wide range of NLP tasks, such as question-answering and language inference, without substantial task-specific architecture modifications [24].

2.2.4 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative probabilistic model that enables sets of observations to be described by unobserved (latent) groups which explain why some parts of the data are similar. Figure 2.9 presents an example of one of the main applications of LDA in Natural Language Processing: the observation of topics from a collection of text corpora [16].

2.3 Cryptocurrency Concepts

This section is organized into two subsections. The first subsection explains blockchain, the technology that utilizes cryptocurrency. The second subsection describes what cryptocurrencies are and how they can be utilized and traded.

	Word 0	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10	Word 11	Word 12	Word 13	Word 14
Topic 0	people	think	mean	state	http	read	number	follow	twitter	fund	send	want	include	group	intelligence
Topic 1	year	market	money	buy	pandemic	business	time	tell	crisis	economy	company	stop	lose	help	law
Topic 2	come	use	want	pay	check	write	trade	build	problem	datum	blockchain	support	know	plan	create
Topic 3	lot	feel	let	agree	think	story	expect	account	space	coin	exchange	line	investment	user	fail
Topic 4	day	com	thing	talk	today	watch	learn	post	share	video	trading	deal	use	sense	miss
Topic 5	time	bitcoin	point	price	love	health	idea	man	care	rate	issue	hope	oil	sell	level
Topic 6	make	know	thank	people	change	month	happen	death	tweet	lockdown	spread	hold	leave	community	lead
Topic 7	government	life	test	question	live	virus	hit	hear	news	hour	join	set	experience	cause	response
Topic 8	work	week	people	world	make	home	mask	end	start	report	stay	place	industry	friend	economist
Topic 9	say	look	way	case	try	run	country	ask	thing	think	break	term	year	podcast	believe

Figure 2.9: LDA Application Example. This table shows the output of the LDA algorithm which includes ten topics and the fifteen most relevant words in each topic. The LDA algorithm was applied to the Cryptocurrency Twitter Dataset that was collected for this thesis.

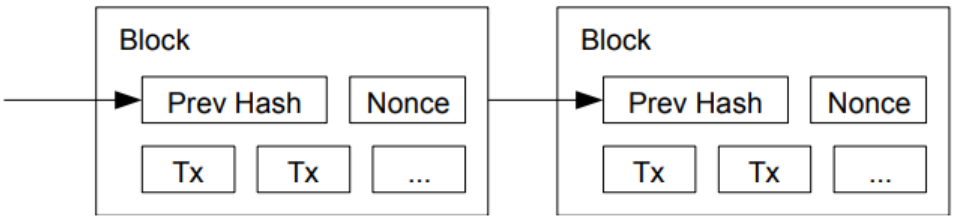


Figure 2.10: Blockchain Structure [52].

2.3.1 Blockchain

Blockchain is a decentralized logging transaction network that is sustained by a peer-to-peer validation system. "Block" is a conglomeration of digital information, such as transaction value, recipient identifier, sender identifier, and transaction time, while the "chain" is how those blocks are interconnected [58]. Once a block is added to the end of the chain it receives a unique identifier hash and contains a reference to the previous block's hash. Therefore, it is comparable to a linked list data structure (Figure 2.10).

This system is decentralized because it is not deployed to one single location, but is rather hosted on thousands of computers [18], also known as blockchain nodes. The Bitcoin blockchain is open source, therefore anyone can create a new node of the network and contribute to its operations. The more nodes on the network the more secure the whole system becomes.

The BTC blockchain is secure due to the complexity and the amount of resources it would take

to misuse it. If someone attempts to alter a transaction history of block B_n, that would require a recalculation of block's B_n hash and an update of all of the hashes of the blocks that come after that ($B_{n+1}, B_{n+2}, ..., B_{lastBlock}$). Furthermore, all of those hash modifications would need to be updated for all of the blockchain nodes (copies) simultaneously, which requires an enormous amount of resources and computational power. Contributors, those that have a node locally deployed, are rather financially motivated to work towards helping the blockchain functionality by mining blocks [52], i.e., by trying to resolve the hashing algorithm of a particular block so it can be added to the chain. Currently, as of October 2020, the reward for mining one block is 6.25 BTC [42], which is worth about $68,750 USD.

Blockchains can be thought of as an evolution of traditional centralized databases by proposing a paradigm that is far more secure and trustworthy. From a social-economical and practical perspective, blockchains are revolutionizing the exchange of value between two parties by removing intermediate entities, such as banks, from this process. This is done by making the exchange of value between parties a much faster process that is always available.

2.3.2 Cryptocurrency

Cryptocurrency is a digital asset that was designed to represent an exchange of value between two parties, like any other fiat currency such as the dollar. Cryptocurrencies are exchanged through their specific blockchain. For example, Bitcoin (BTC) can only be exchanged through its own blockchain, while Ethereum (ETH) cannot be sent through BTC's blockchain. Consequently, each cryptocurrency has its own wallet. Unlike fiat currencies, cryptocurrencies cannot be fabricated when they reach their pre-established maximum supply. There will only be 21 million BTC in the world, and if its demand keeps increasing, the price will also increase due to the limited supply. Cryptocurrencies have a deflationary property, since its purchase power increases over time.

Like fiat currencies, it is also possible to invest and trade with cryptocurrencies. There are

Ticker Symbol	Cryptocurrency	Market Cap	Price	Circulation Supply
BTC	Bitcoin	$218B	$11,780	18,522,075
ETH	Ethereum	$43B	$378.62	113,080,246
Tether	Tether	$16B	$1.00	15,857,387,815
XRP	XRP	$11B	$0.24	45,248,061,374
BCH	Bitcoin Cash	$5B	$249.17	18,549,356
BNB	Binance Coin	$4B	$29.89	144,406,561
LINK	Chainlink	$4B	$10.77	388,509,556
DOT	Polkadot	$3B	$4.06	852,647,705
ADA	Cardano	$3B	$0.10	31,112,484,646
LTC	Litecoin	$3B	$48.17	65,705,853

Table 2.1: Top 10 Cryptocurrencies Information (Current as of October 2020).

Figure 2.11: Cryptocurrencies Logos.

crypto-specific trading platforms such as Binance[4] and Coinbase[5], but trading is also possible through stock exchange platforms such as Robinhood[6], eToro[7], and others. Cryptocurrencies are not only identified by their names, but like stock indexes, they have their own letter code (ticker symbol). Table 2.1 contains the top ten cryptocurrencies by market capitalization (market cap) [8] with their respective ticker symbol. Cryptocurrencies are not only distinguished by their names and ticker symbols but like fiat currencies, they also have a logo branding (Figure 2.11).

[4]https://www.binance.com/

[5]https://www.coinbase.com/

[6]https://robinhood.com/us/en/

[7]https://www.etoro.com/

[8]Market capitalization is the value that represents the price of each unit, such as cryptocurrency, times its circulation supply.

CHAPTER 3

RELATED WORK

This chapter presents previous work related to the contributions of this thesis and is divided into five main sections. The first section is an overview of online discourse and its effects on public opinion publications. The following section is about similar work studying Twitter sentiment for stock market prediction. The third section goes over the optimal historical data collection, while the next section introduces the few works concerning cryptocurrency price prediction. Section 3.5 focuses on exemplifying several relevant work on framing theory in microblogs. Finally, this chapter concludes with a discussion of the novel contributions of this thesis.

3.1 Online Discourse and Effects on Public Opinion

Modeling social media microblogs, specifically Twitter, to show connections between online discourse and its effects on public opinion has been widely studied in NLP [8, 32, 64, 68, 69, 60] and the social sciences [12, 17, 31, 49, 37]. The study on the examination of framing effects on the Vancouver riots [17] demonstrates how Twitter is not only a source of information, but also a way of shaping people's opinions and their cultural perceptions. Most of the work in online discourse and its effects in public opinion are related to cultural and political events, such as American politics [19, 39], and the 2011 Egyptian protests [31].

Currently, there is no work analysing online discourse and its effect on public opinion for stock market trends, let alone for cryptocurrency trading movements.

3.2 Twitter Sentiment for Stock Market Prediction

There are many works on Twitter sentiment analysis, but closest to this thesis are those concerning the use of Twitter sentiment for stock market predictions [43, 57, 63, 22].

Derakhshan & Beigy, in their work, "Sentiment Analysis on Social Media for Stock Price Movement Prediction", proposed a new opinion mining model based on LDA and Part-of-Speech

(POS) features. Their work aims to predict American and Persian stock market movements with their LDA-POS graphical model. This model is heavily based on sentiment analysis and only focuses on predicting the up and down trends of stock prices, but does not account for when there is no price change or movement.

3.3 Optimal Historical Data Collection

Walczak has focused on both how much input is necessary for optimal time series modeling, and has outlined the adequate amount of historical data required to produce the best performing neural network models for financial forecasting [67]. According to Walczak's work, financial time series predictions require two years of training data as the optimal time period to forecast future fiat currency exchanges. Different from this work, this thesis focuses on predicting cryptocurrency investment *actions*, instead of fiat currency prices, by extracting patterns from historical *tweets* rather than stock values and indices.

3.4 Cryptocurrency Price Prediction

There are relatively few works concerning cryptocurrency analysis and prediction. Of these, a majority use social media sentiment [36, 48], volume of tweets [66], or both [7] as the main feature for prediction. Furthermore, the prediction tasks are typically to predict cryptocurrency prices or whether the prices will rise or fall.

Li's sentiment-based prediction model [48] is the first to demonstrate that social media microblogs, such as Twitter, can be used for predicting price movements in such a speculative market as smaller cryptocurrencies, also known as alt-coins. This work, however, only focuses on analyzing the ZClassic alt-coin market. The model is an Extreme Gradient Boosting Tree (XGBoost) that utilizes Twitter sentiment and trading volumes to predict price fluctuations.

Abraham's work [7] is also based on an XGBoost model to predict Bitcoin price fluctuation. This model, like Li et al., also utilizes sentiment and cryptocurrency prices as features. Abraham's work differentiates from Li's [48] by creating a real-time architecture and predicting time fluctuations

for a different cryptocurrency: Bitcoin.

3.5 Framing Theory in Microblogs

Previous works have shown the effectiveness of using frames to predict various social sciences phenomena, such as political framing of Twitter discourse, congressional speeches, and news coverage of current events [10, 13, 19, 26, 37, 39, 65, 25].

Card's contribution to the NLP community is the development of a human annotated media framing corpus [19] based on a well-developed guideline. The Media Frames Corpus consists of thousands of news articles and focuses exclusively on how three policy issues (immigration, smoking, and same-sex marriage), are framed in the media.

Johnson's work [39] goes a step further by proposing a collection of weakly unsupervised models to predict frames in the tweets of politicians. This work stands out by combining lexical features of tweets and network-based behavioral features that results in a substantial improvement over a lexical baseline.

Most recently, Field [25] focused on identifying and analysing media manipulation utilizing cross-lingual projection of framing annotation to prove political agendas such as distracting Russian citizens away from the Russian economic crisis by bringing to their attention negative news events in the United States.

Political framing on Twitter has also been studied by looking at one issue at a time, such as climate change. Jang's work aims to understand how climate change frames are incorporated into everyday conversations, i.e., who uses "global warming" versus "climate change", and from what state and countries these people are from [37].

3.6 Novel Contributions

Sentiment is known to be difficult to predict on Twitter. Furthermore, the volume of tweets can be falsely inflated by bots reporting currency prices, but not contributing to the discourse. Therefore, instead of sentiment or tweet volume, this thesis aims to use the language directly

extracted from tweets, their context, and features representing the social network behavior for a buy, sell, or hold investment action prediction. Furthermore, this work is the first to explore framing in the cryptocurrency domain, as well as in economics. In order to extract the frames, Deep Embedding Clustering with Bag of Words as features were used.

Despite this coverage, at the time of writing this thesis there are no Natural Language Processing publications studying the role of framing in economics, specifically concerning Wall Street stocks or cryptocurrency day trading, or associated correlations with the current pandemic. This work represents a first step in understanding how framing can reveal insights into cryptocurrency day trading actions.

CHAPTER 4

DATA ANNOTATION

This chapter presents the tweet collection and preprocessing steps, as well as the collection of historical Bitcoin (BTC) transaction prices, used to construct the datasets. Section 4.1 concentrates on the Twitter data collection, its categories, and volume distribution. Section 4.2 focuses on BTC historical price collection, while Section 4.3 aims to describe the tweets preprocessing steps. Finally, Section 4.4 describes how the tweets were annotated for use in the weakly-supervised day trading behavior prediction model. The non-annotated version of these tweets are used in the clustering models.

4.1 Twitter Data Collection

For this work, tweets related to cryptocurrency were collected including Bitcoin and other coin types such as Ethereum (ETH) and XRP. Rather than collect based on hashtag or keywords alone, the search was narrowed to specific time frames and user accounts. Tweets were scraped from January 2017, when Bitcoin surpassed $1,000 per coin, to its last all time high price in November 2013, and then again until March 2020. This timeline covers times of frequent changes in cryptocurrency trading and adheres to the finding that an optimal dataset for financial time series prediction consists of information from the past two years [67]. These tweets form the *Pre-COVID* (before the pandemic) Dataset.

Within these time frames, three types of user accounts were identified for tweet collection to maximize presence of discourse for analysis and minimize tweet noise. These include influential cryptocurrency Twitter accounts, or *influencers*, which are well known as sources for investment information and thus should provide features for message propagation. This category also includes users who frequently tweet about cryptocurrency and have at least ten thousand followers. Similarly, *media* accounts from traditional or online news sources, such as *@CNNBusiness* and *@BitcoinMagazine*, are used. Lastly, there are the *company* accounts, such as *@IBMBlockchain*

Quantity	Influencers	Media	Company
$10,000 - 99,999$	45	-	13
$100,000 - 499,999$	24	2	5
$500,000 - 999,999$	2	2	1
$\geq 1,000,000$	-	5	-

Table 4.1: Quantity of Followers Per User Account Type. Each row represents the number of user account types (columns) that have that quantity of followers who are actively tweeting about cryptocurrency.

Dataset	Influencers	Media	Company
Before Pandemic (Pre-COVID)	136,637	128,041	110,846
During Pandemic (COVID)	48,254	24,014	36,233

Table 4.2: Quantity of Unique Tweets Per User Account Type. Each row represents the number of tweets of each account type (columns) appearing in each dataset.

and *@BitPay*. By narrowing down the search to these well-known and highly followed accounts, a lot of Twitter noise was removed, e.g., dropping tweets that mention cryptocurrency but do not relate to its purchase or trends.

Table 4.1 presents the distribution of followers for accounts collected from the different types of accounts mentioned above. Column one lists the quantity of followers, divided into four groups. The remaining columns indicate how many of the influencer, media, and company accounts have the different number of followers. From this table, it is clear that the majority of tweet activity comes from influencer accounts that have between 10,000 and 499,999 followers. There are fewer media accounts, however, these accounts have much broader reach. For example, *@nytimes* reaches up to 46.6 million people when tweeting about cryptocurrencies.

Using the same accounts, additional cryptocurrency tweets were collected which occurred during the COVID-19 pandemic time frame: from February 2020 until June 2020 [1]. These tweets comprise our *COVID* (during the pandemic) Dataset. The total amount of tweets collected for both datasets is 530,911, where 407,396 belong to the Pre-COVID Dataset and 123,515 belong to the COVID Dataset. Table 4.2 summarizes the amount of unique tweets per account type that appear in the two dataset collections.

[1] Though the pandemic continued after this time frame, this is when the last tweets were collected.

Date	Open*	High	Low	Close**	Volume	Market Cap
23-Feb-20	9663.32	9937.4	9657.79	9924.52	4.12E+10	1.81E+11
19-Jan-20	8941.45	9164.36	8620.08	8706.25	3.42E+10	1.58E+11
25-Nov-19	7039.98	7319.86	6617.17	7146.13	4.27E+10	1.29E+11
1-Jun-19	8573.84	8625.6	8481.58	8564.02	2.25E+10	1.52E+11
2-Dec-18	4200.73	4301.52	4110.98	4139.88	5.26E+09	7.21E+10
9-May-18	9223.73	9374.76	9031.62	9325.18	7.23E+09	1.59E+11
31-Oct-17	6132.02	6470.43	6103.33	6468.4	2.31E+09	1.08E+11
6-May-17	1556.81	1578.8	1542.5	1578.8	5.83E+08	2.58E+10

Table 4.3: Sample of BTC Historical Price Dataset.

4.2 BTC Historical Price Data Collection

In addition to cryptocurrency related tweets, historical transaction prices of Bitcoin were collected from CoinMarketCap [2]. This BTC Historical Price Dataset contains the following information: the opening price of Bitcoin (Open), the highest price (High), the lowest price (Low), and the closing price (Close) of Bitcoin on that particular day (Table 4.3). This dataset also includes the date and the dollar volume of BTC traded that day.

4.3 Preprocessing

Before processing, a total of 407,396 tweets with meta-information, including number of replies, number of retweets, and the date, were collected. Preprocessing consisted of three main steps. First, all tweets were standardized by controlling for capitalization, applying stemming, and removing URLs, white space noise, and stop words. Second, irrelevant tweets were removed by filtering for the presence of cryptocurrency-based keywords or hashtags (e.g., Bitcoin, BTC, Ethereum, crypto, cryptocurrency, blockchain, XRP, altcoin, etc.) reducing the dataset to 64,685 tweets.

The collected tweets were labeled as *buy*, *sell*, or *hold* depending on their price change from one day to another. In order to determine the minimum percentage gain to label certain tweets as *sell*, the BTC volatility baseline had to be determined and compared to the regular stock market. Volatility is the degree of variation of a trading price series over time, normally measured by the

[2] https://coinmarketcap.com/

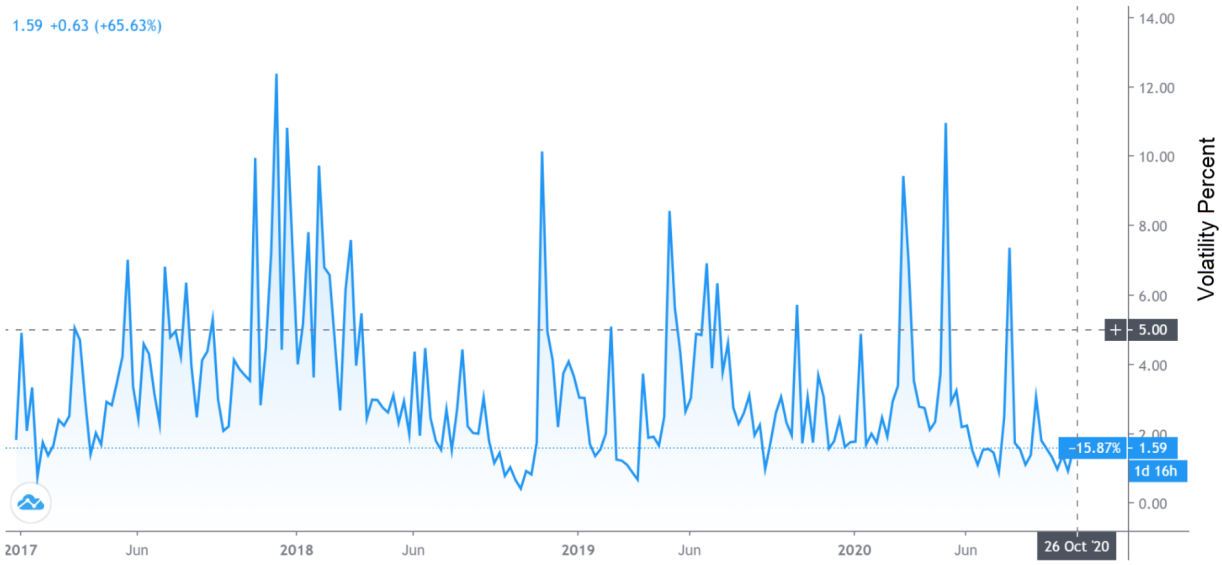


Figure 4.1: BTC Volatility from 2017 to 2020 [2].

standard deviation or variance of logarithmic returns. Volatility is usually associated with big swings of trading price in either direction [44]. The average volatility of regular stock day trading is 3.3% which is a high value according to Kyröläinen. However, BTC volatility is much higher than the stock market (as shown in Figure 4.1), especially between 2017 and 2018 when it was around 8% [59]. Between 2019 and 2020 it was lower at 4.66% [1], however this is still higher than the average volatility for the stock market.

Based on BTC and day trading volatility information, for this work 5% of price movement was chosen to focus on understanding the influence of tweets during the highest peaks of volatility. Therefore, tweets that corresponded to days with at least a 5% increase or decrease of BTC price were retained. The price movement of 5% was calculated by taking the price difference between the current day's closing price and the past day's closing price. After processing, a total of 18,900 filtered tweets were used for experiments regarding the day trading movement prediction.

For the frames clustering and experiments, an additional 123,515 tweets were collected during the pandemic time span. Preprocessing for these tweets consisted of removing: duplicate tweets, English stop words, and references to other users, emails, or website links.

Date	Tweet	Reply Count	Retweet Count	Account Type
5/1/17	at hash rate of 4 000 000 th bitcoin is secured by over half-billion dollars of hardware	3	45	influencers
5/8/17	bitcoin investment trust ups its proposed ipo but approval is still in question bitcoin investing etf fintech	1	14	media
8/11/17	can bitcoin disrupt the payment processing industry	13	98	media
9/15/17	cnbc bitcoin fans fire back at jamie dimon after fraud comment	36	116	most followed
12/17/18	goldman sachs has been criminally charged by Malaysian officials for their participation in the 1mdb scandal long bitcoin short the bankers	35	239	influencers
3/23/20	bearish momentum keeps prevailing btc	2	3	most followed

Table 4.4: Sample of the Day Trading Tweets Dataset After Pre-processing.

4.4 Annotation

In order to create an annotated dataset for training and testing a weakly-supervised day trading prediction model, the price information in the BTC Historical Price Dataset (Section 4.2) was used. With this information, a *momentum* metric that represents the fluctuation of cryptocurrency costs on a given day was defined:

$$momentum = \frac{Price_{close} - Price_{open}}{Price_{open}} \tag{4.1}$$

If the momentum on a given day increases or decreases by five percent on the following day, then the tweets of that given day are labeled as *buy* or *sell*, respectively. If there is less than five percent change, these tweets are neutral in terms of buying or selling, and are therefore labeled as *hold*, to represent that an investor should take no action with their cryptocurrency. The annotation was automated with a script that cross referenced the date of the tweet with the BTC Historical Price Dataset.

Recall that the first goal of this work is to predict whether an investor should buy, sell, or

hold their cryptocurrency based on the tweets discussing cryptocurrency that day. Given the high quantity of tweets and highly dynamic language of Twitter, as well as the subjectivity of choosing to buy, sell, or hold, the momentum metric is chosen as a weak form of supervision for investment actions.

To further strengthen the hypothesis that trading prediction is a challenging task, two annotators, with different investment experience backgrounds, were asked to label (*buy*, *sell*, or *hold*) a randomly generated subset of the Pre-COVID dataset based on the the tweet content, tweet author, and BTC price percentage fluctuation from the previous day. The reduced dataset for manual annotation has 798 different tweets. There are 114 different days represented in the dataset with 7 distinct tweets per day.

The annotators were asked to perform three different experiments. First, they were asked to label the tweets based on their content. After labeling all the tweets for a particular day individually, they were asked to give an overall label for that particular day based on all their individual tweet annotations. Finally, they were asked to give another overall annotation for a particular day with the additional information about the BTC price percentage change from the previous day.

For the annotation experiments, both annotators had different levels of experience in both investing and trading stocks and cryptocurrencies. One of the annotators was an inexperienced investor, who has never bought or sold cryptocurrencies or stocks. Furthermore, the inexperienced annotator has heard of Bitcoin and blockchain, but has limited knowledge on how blockchains and cryptocurrency work. Furthermore, this annotator was unfamiliar with what tools and applications are needed to start investing in cryptocurrencies. The second annotator is an experienced investor that has been investing and following the stock market for the past 5 years, and in the past 2 years has been investing in cryptocurrencies. However, the experienced investor has a long term strategy, which means this annotator does not practice day trading. The second annotator also has a very broad knowledge about cryptocurrencies, blockchain, and investment tools.

Table 4.5 reports the results of the first experiment, labeling tweets based on their content, for both annotators. The true labels are the ones generated by the momentum equation (Equation 4.1).

Label	Inexperienced Annotator Precision	Experienced Annotator Precision
Sell	32%	17%
Buy	32%	33%
Hold	38%	33%

Table 4.5: Annotation Experiment One. Tweet by tweet annotation precision from an annotator that has never invested and an experienced long term investor.

Label	Inexperienced Annotator Precision	Experienced Annotator Precision
Sell	17%	20%
Buy	28%	36%
Hold	36%	31%

Table 4.6: Annotation Experiment Two. Overall day-based tweet annotation from an inexperienced and experienced investor.

Label	Inexperienced Annotator Precision	Experienced Annotator Precision
Sell	50%	0%
Buy	30%	0%
Hold	34%	53%

Table 4.7: Annotation Experiment Three. Overall day annotation based on tweet content and BTC price percentage change from the previous day.

The majority of the results are close to random guessing (33%), besides the 17% precision of the *sell* label generated by the experienced annotator.

For the second experiment, where the annotators were asked to give an overall label for the day, both annotators performed significantly below random guessing, as illustrated in Table 4.6, where the expected label was *sell*.

In the last experiment, annotators had to take into consideration the price movement from the previous day to decide on what trading action, *buy*, *sell*, or *hold*, to take. The annotators have very contrasting results, as shown in Table 4.7. The inexperienced annotator outperforms random guessing by over 15%. This is likely because their strategy was to sell when prompted with a

strongly worded tweet combined with big BTC price drops. Furthermore, the other annotator did not perform well because a long term investing strategy was applied to a day trading application. The experienced trader invests in BTC with the goal to profit from it in the next 20 years, therefore, when there is a drop in the price this investor sees it as an opportunity to buy more, while for a day trading strategy, selling when the price is going down is one of the mechanisms to reduce losses in the short term.

The results of these annotation experiments illustrate that day trading is a non-trivial task for people that do not have any prior trading and investing experience, as well as for those who do have such experience. Given the variance in labeling via human annotators, the momentum metric was used to generate weakly-supervised labels for the day trading prediction experiments in this work.

CHAPTER 5

MODELING AND FEATURE ENGINEERING

In this chapter, the two modeling approaches of this thesis are described. The first one is a weakly supervised model to predict BTC day trading behaviors, while the second approach is an unsupervised model to extract discourse framing clusters from cryptocurrency tweets. The features associated with each experimental model are also discussed in this section. These features represent both aspects of the social network nature of Twitter and the actual language and context of the tweets.

5.1 Day Trading Behavior Prediction

The day trading behavior prediction model is designed to predict a *buy*, *sell*, or *hold* label given tweets coming from the media, known people in the cryptocurrency space (*influencers*), and highly followed cryptocurrencies accounts. The objective of this task is to guide investors on trading decisions based on the tweet labeling.

5.1.1 Day Trading Model

For the day trading prediction model experiments a combination of features and models were executed to both determine the most relevant features and the best model for the task. Naive Bayes, with Bag-of-Words (BOW) features, was used for the baseline model. During the experiments, a Conditional Random Field (CRF) and XGBoost were tested with a set of different features. However, those models either did not converge or yielded results very close to random guessing. Therefore, these models were deemed not appropriate for the prediction task as is described in Chapter 6. Ultimately, Random Forest, RNN, and LSTM models were chosen for further development, and their experiments resulted in final accuracies above 85%. The RNN with three layers described in Section 6.1 is the best performing model for this task.

Topic	Words
Develop Network	bitcoin, price, index, usd, year, need, develop, value, today, investor, question, bch, offer, tech, network
Sell Bitcoin	buy, peopl, time, money, support, use, think, day, ethereum, btc, sell, know, month, bitcoin, market
Blockchain	blockchain, make, market, look, invest, trade, say, want, use, fintech, come, chain, pay, learn, ripple

Table 5.1: LDA Topics and Their Corresponding Words.

5.1.2 Day Trading Model Features

Social network features are extracted directly from the meta-information of the cryptocurrency tweets. This includes the number of retweets and the number of replies. During the experiments, it was observed that the number of retweets provided some information gain when weighting the tweet feature representation. The type of user account, either influencer, media, or company, that posted the tweet is also used as a feature.

In addition to these features, features directly related to the language of the tweet were used to determine how much additional features would contribute to the final model. First, an LDA topic model [38] was implemented. From this, the top three topics were extracted and the presence of the topic in a given tweet was used as a feature. Table 5.1 shows the top three LDA topics that were used, *Develop Network*, *Sell Bitcoin*, and *Blockchain*, and their respective words. The LDA topic distribution was also extracted from the dataset to understand the patterns between the topics and each trading category. Figure 5.1 shows that the most relevant topic for the subset of days that are labeled as *buy* is *Develop Network*, while the topic with lower frequency is *Sell Bitcoin*. Exactly the opposite happens when observing the distribution of topics, as shown in Figure 5.2, for the *sell* category. The most relevant topic for the subset of *sell* tweets is *Sell Bitcoin*, and the least relevant topic is *Develop Network*. However, the distribution of topics for tweets that are labeled *hold* is very similar to the *buy* distribution, as shown in Figure 5.3.

Next, the tweets were transformed into 768 language features using DistilBERT [62], a contextual embedding modeling framework. Typically NLP works represent tweets as features using the original BERT model or one of its variants. During the initial experiments for this thesis,

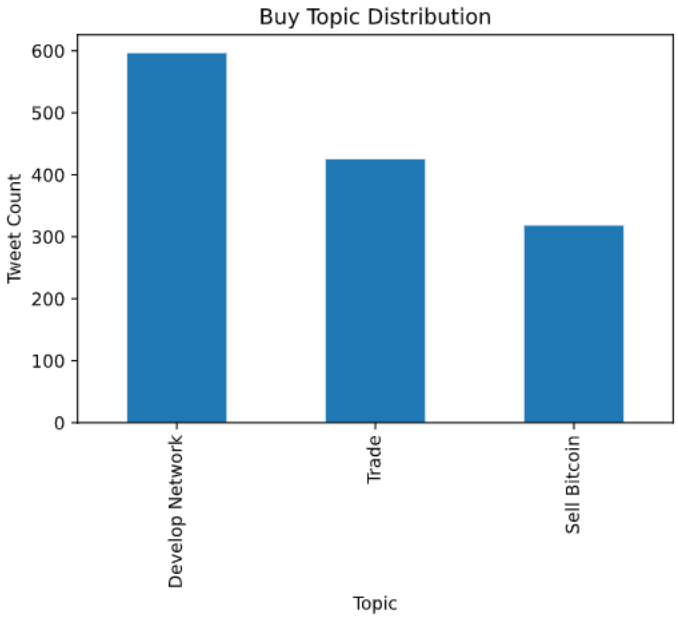


Figure 5.1: LDA Topic Distribution for *Buy* Tweets.

DistilBERT had a 0.6% better performance than BERT, and was therefore used for language feature representation in the model. All of the tweets were concatenated according to their momentum label and for each group (*buy, sell, hold*), DistilBERT was used to extract high-quality language features to represent each of the three tweet groups.

In addition to these DistilBERT-based representations, the cosine similarity was calculated for each tweet of the three tweet group representations above. Further, the match between a tweet and group with the highest cosine similarity was selected to be used as a feature for that tweet. More concretely, each tweet is compared to the DistilBERT representation of the buy, sell, and hold concatenated tweet groups and the highest similarity group is chosen to be used as a feature.

The most relevant features identified from the Random Forest (Figure 5.4) were extracted and plugged into the Recurrent Neural Network. However, they did not outperform the recurrent network that was using DistilBERT representation as features. Details of the results of these models are further discussed in Chapter 6.

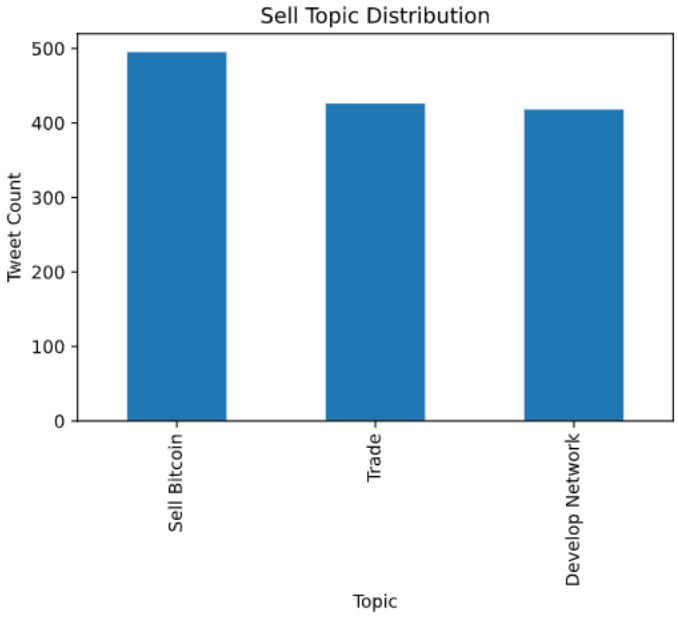


Figure 5.2: LDA Topic Distribution for *Sell* Tweets.

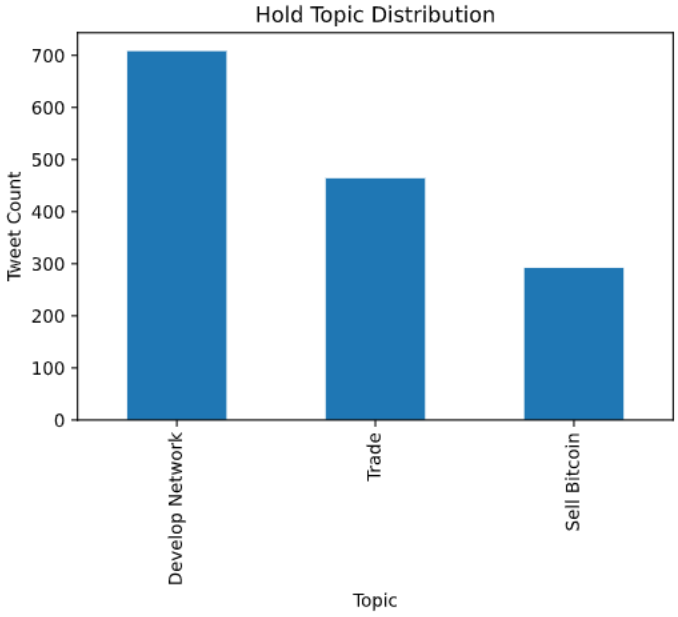


Figure 5.3: LDA Topic Distribution for *Hold* Tweets.

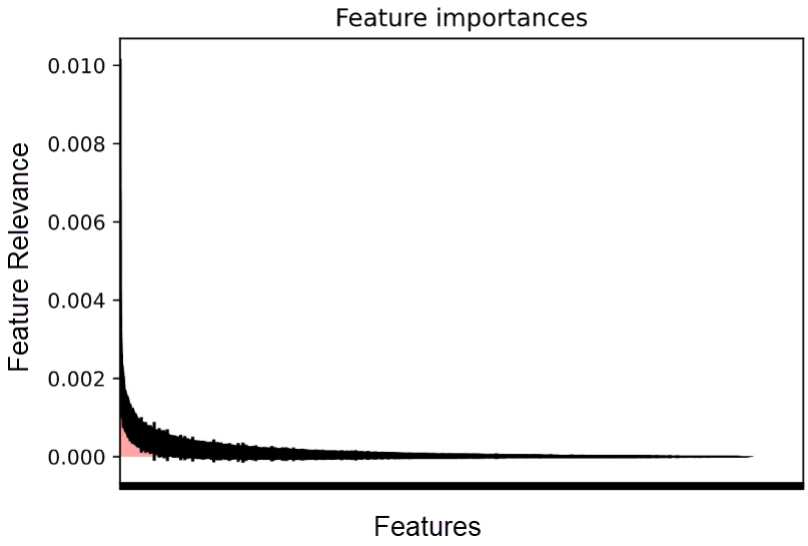


Figure 5.4: Random Forest Feature Relevance Distribution.

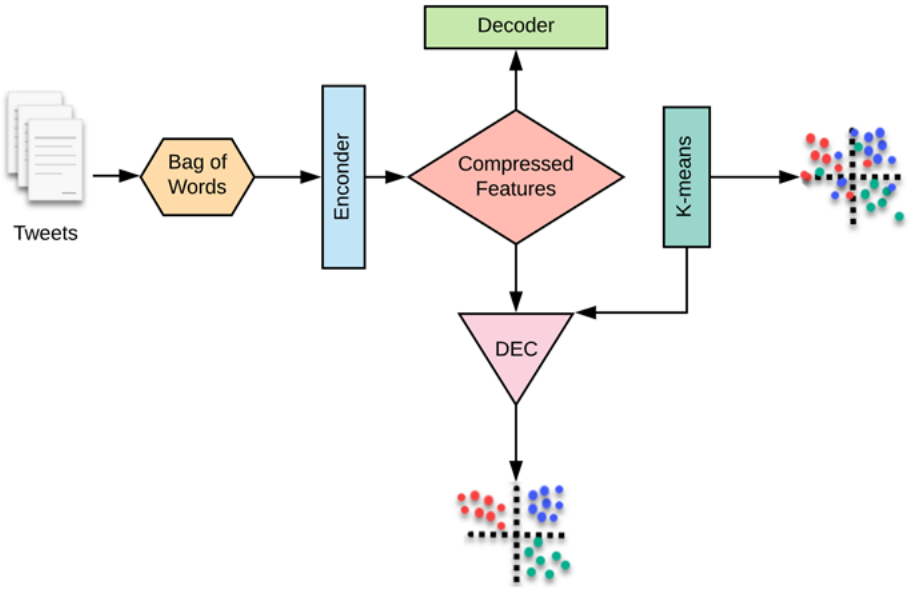


Figure 5.5: Autoencoder and Deep Embedded Clustering (DEC) Pipeline. DEC clusters the data by simultaneously learning a set of k cluster centers in the transformed feature space from the autoencoder.

5.2 Discourse Framing Clustering

The Discourse Framing Clustering model was implemented to extract the initial frames for the Cryptocurrency Tweets Dataset.

5.2.1 Discourse Framing Model

From an NLP perspective, frames are nuanced, latent abstractions of a discussion. The hypothesis is that *how a topic is discussed, or framed,* could be identified in an unsupervised manner by analyzing how the tweet content clusters together. To extract the clusters which represent such frames, two modeling approaches were implemented. First, a basic k-means clustering approach was chosen as the baseline model. Second, an unsupervised Deep Embedded Clustering (DEC) approach [70], which combines both an autoencoder and k-means clustering to achieve a more precise separation, was used. As shown in Figure 5.5, DEC simultaneously learns feature representations and cluster assignments.

Features. The features used for the basic k-means clustering and DEC models were a sparse representation of the word count for each tweet. Both BOW and TF-IDF features were used as input to the k-means clustering model and autoencoder of the DEC pipeline. TF-IDF stands for *term frequency–inverse document frequency*, and it is a numerical statistic that represents the importance of a word to a sentence or document within a corpus [56]. Both BOW and TF-IDF features were built on top of the unfiltered dataset, meaning that besides duplicated entries, no tweets were removed.

CHAPTER 6

EXPERIMENTAL RESULTS

In this chapter, the experimental setup, trials, and analysis of modeling results are presented for both the day trading behavior prediction and discourse framing prediction models. First, this chapter covers the trial approach for the day trading behavior prediction, including the justification to choose and pursue the work with an RNN instead of the CRF model. The reasoning behind focusing on language feature representations for the day trading behavior prediction task is also discussed. Next, Section 6.2 shifts the focus of this chapter to discuss the experimental findings for the discourse framing prediction.

6.1 Day Trading Behavior Prediction

The supervised experiments were conducted using five-fold cross-validation with random shuffling and an 80% training and 20% testing split. For the neural networks, 50 epochs were chosen because the dropout after each layer was 0.001.

Prior to focusing on a subset of models, experiments were conducted using CRF and XGBoost. The main challenge with the CRF model was the lack of convergence when using the tweet content representation in the form of unigrams as a feature with the intent to predict *buy*, *sell*, or *hold* labels as the prediction task (Table 6.2). In order to reduce the dimensionality of the task for the CRF, which would facilitate convergence, the experiment was modified to try to predict LDA topics instead of *buy*, *sell*, or *hold* as shown in Table 6.1. Additionally, the labels *buy*, *sell*, or *hold* are now input to the model as features. However, random guessing was still close to the CRF accuracy, while the Random Forest on the same task performed 9% more accurately than CRF and 14.67% better than random guessing.

XGBoost was able to predict, with an accuracy of 45%, the buy, sell, or neutral class for the original task better than random guessing, nevertheless it did not outperform the Random Forest which achieved 87.09% accuracy on the same assignment as shown in Table 6.2. In order to

Model	Predict	Features	Accuracy	Random Guessing
RF	3 LDA Topics	Buy, sell, hold, No. of Replies, No. of Retweets, Category	48%	33.33%
CRF	3 LDA Topics	Buy, sell, hold, No. of Replies, No. of Retweets, Category	37%	33.33%

Table 6.1: RF and CRF Comparison. Experimental results with Conditional Random Fields (CRF) and Random Forest (RF) when predicting 3 LDA topics based on the buy label, sell label, hold label, number of replies and retweets, and the category as features.

Model	Predict	Features	Accuracy	Random Guessing
RF	Buy, sell, or hold	Unigram of tweets	63%	33%
XGBoost	Buy, sell, or hold	Unigram of tweets	45%	33%

Table 6.2: RF and XGBoost Comparison. Experimental results with XGBoost and Random Forest (RF).

understand the impact of the features on the RF model, the experiment of Table 6.3 shows that the model that takes in the additional LDA topic as a language feature performs significantly better than the model that does not take in any language feature representation. The language-based feature is the most significant feature of the model, which can be inferred by looking at the RF model performance in Table 6.2. With these initial experimental results, the CRF and XGBoost studies were dropped and further development was dedicated to analyzing neural network performance on this novel task.

Table 6.4 shows the results of using the following models: Naive Bayes, Random Forest, Recurrent Neural Network, and an LSTM. Both the RNN and LSTM use three dense layers. The columns of Table 6.4 correspond to the tweet feature representations used with each model: a baseline where tweets are represented as Bag-of-Words (BOW) and DistilBERT as described in Section 5.1. Ablation studies revealed that the most informative features for prediction were

Model	Predict	Features	Accuracy	Random Guessing
RF	Buy or sell	No. or Replies, No. of Retweets, Category	41%	50%
RF	Buy or sell	No. or Replies, No. of Retweets, Category, LDA topics	59%	50%

Table 6.3: Experimental Results with Random Forest (RF). One experiment used LDA topics as features while the other did not.

Model	BOW	DistilBERT
Naive Bayes	49.72%	61.58%
Random Forest	63.81%	87.09%
RNN	33.67%	88.78%
LSTM	31.57%	88.18%

Table 6.4: Day Trading Prediction Results. The columns represent the accuracy of each model when using either a Bag-of-Words (BOW) or DistilBERT [61] representation of the tweets as features.

the language features, specifically the combination of DistilBERT representations with cosine similarity.

From Table 6.4, it is possible to observe that using an RNN with DistilBERT has the highest accuracy of 88.78% across all three classes. Predicting day trading behavior, i.e., whether to buy or sell stock, is a complicated task, especially in a volatile asset such as cryptocurrency. By carefully preprocessing the dataset and using DistilBERT for tweet language representation as a feature, both the LSTM and RNN architectures were able to yield high accuracy on this challenging task.

6.2 Discourse Framing Prediction

Unsupervised clustering experiments were conducted using: (1) a basic k-means clustering algorithm and (2) deep clustering with autoenconders (DEC) [30] as described in Section 5.2. The encoder outputs were used as inputs to the deep clustering layer, and the k-means center clusters were used as initial weights for the deep clustering model. The tweets were randomly shuffled for

Cluster	Tweet Content
Politics	I expect crypto currencies will become "normalized" in the Indian market over time. I hope the reactionary govt actions are shortlived
Politics	Already the case in a number of countries where it is banned, yet has increased in use. Example, Venezuela.
Politics	The guy from Venezuela who wrote the post is sharing why Bitcoin was working and the banks weren't when the power was out...
Politics	Will 2018 be the year for blockchain for government?

Table 6.5: Example Tweets Per Cluster Type in the Pre-COVID Dataset.

training. The autoencoder ran for 100 epochs, achieving an accuracy of 99.99% with both training and validation loss on the order of 5.5453e-04 without overfitting.

Initially the experiment was executed with 32 clusters because 32 is the default number of features that get compressed by the autoencoder. However, it was observed that several clusters had similar and overlapping topics and keywords. Therefore, the rest of the experiments were conducted with 10 clusters. Figure 6.1 shows the number of tweets that fall into each of the 10 initial clusters for each modeling approach.

Figure 6.1 shows the six predominant clusters identified in the Pre-COVID Dataset by k-means clustering. Using Singular Value Decomposition (SVD) (Figure 6.2) and an analysis of the most frequent words appearing in each cluster, it was possible to extract three main clusters. The first cluster included tweets discussing Bitcoin halving, which refers to the mining capacity of BTC. About every four years or so, the amount of BTC that can be mined decreases by half (halving). With this halving mechanism in place to control the amount of BTC that becomes available in the network over time on top of the demand increases, the price of BTC historically has gone up and remained stable. Therefore, BTC halving is associated with price increase. The second cluster concerns trading and investing cryptocurrency, and the third discusses how trading is affected by politics.

The DEC clustering of the Pre-COVID Dataset (Figure 6.2) identified four main clusters: one discussing halving but with more emphasis on long term store value, one discussing political

effects, and two discussing cryptocurrency trading and applications. This latter cluster splits the *cryptocurrency trading and investing* cluster identified as one large cluster by k-means into two clusters. Chapter 7 provides more analysis of the frames these clusters represent and how they change during the pandemic.

6.2.1 Cluster Verification

Since both clustering approaches operate in an unsupervised setting, an evaluator was tasked with determining how well the clusters represent how cryptocurrency discussions are framed. Given a subset of tweets, the evaluator was asked to label if cryptocurrency was discussed in the tweet with one of the DEC-identified frames using the following guidelines:

- Trading Frame: Does the tweet discuss how or why to buy or sell cryptocurrency?
- Application Frame: Does the tweet emphasize uses of cryptocurrency?
- Store Value Frame: Does the tweet discuss cryptocurrency in terms of long term value?
- Political Frame: Does the tweet put a political spin on cryptocurrency trading actions?

The evaluator's manual annotation was compared to the actual cluster (or frame) the tweet was assigned to by the DEC model. With this evaluation approach, the clustering turned out to be 69.23% accurate. Given the lack of previous work on cryptocurrency framing, this result was compared instead to a previous work that was executed on a tweet dataset labeled for political frames which found an annotator agreement of 73.4% [39]. Next, a chi-squared test was performed to verify the hypothesis that the clusters were dependant on certain words. In order to perform the test, the top word count was collected for each cluster, as well as their count in every other cluster. For example, *coronavirus* was a top word in one of the clusters, therefore the frequency of *coronavirus* was observed and compared in every cluster generated by DEC. The resulting *p-value* was less than 0.05, meaning that the words are highly dependent on the cluster. Therefore, this

Topic	Top Words
Knowledge	know, bitcoin, time, blockchain, market, world, buy, change, people, point, today
Business	year, thank, start, problem, business, write, stop, plan, risk, reason, check
Support	make, think, work, want, day, people, need, use, year, week, support, happen, read
Hold	look, price, money, try, build, econ, think, end, tell, idea, people, term, win, hold

Table 6.6: Pre-COVID Dataset Top 4 LDA Topics and Most Frequent Keywords.

result is reasonable given the unsupervised and novel aspect of this task, as well as the difficulty of determining frames in text and within tweets.

To further support that these clusters could represent how tweets are framed, an LDA topic analysis was performed to ensure that clusters were not finding topics. Table 6.6 shows the top four LDA topics, which are more varied than those extracted for frames (as discussed in more detail in Chapter 7). These topics represent the content of the tweet, e.g., the topic *Hold* represents holding (not buying or selling) cryptocurrency. Frames, however, are fundamentally different and represent *how* someone discusses that topic. A Trading Frame discussing the hold topic can be presented in the form of giving credibility to people that do not sell their cryptocurrency and criticizing those that sell their crypto assets during a crisis, as evidenced by the following tweet: *"Liquidity crisis is happening. Not a big deal long term. Weak hands selling to strong hands right before the halving"*- APompliano.

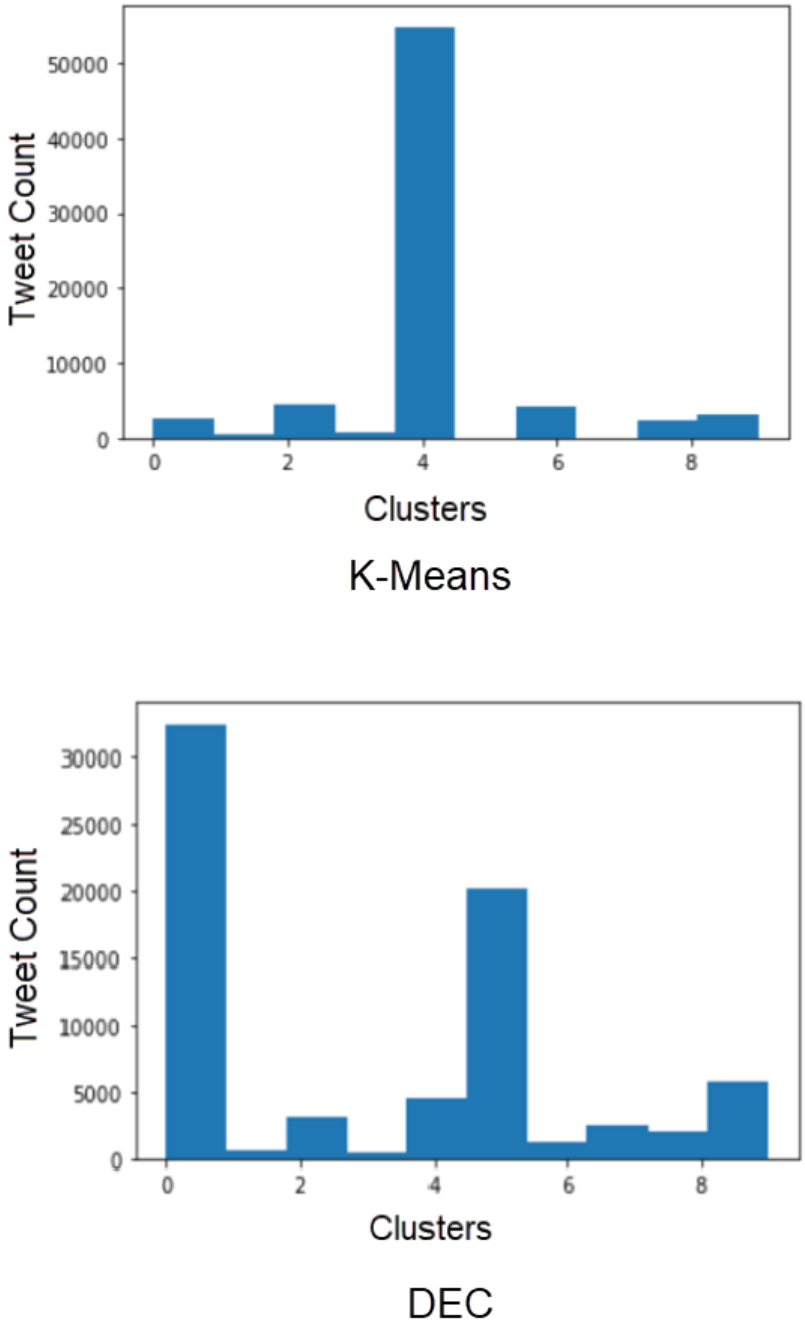
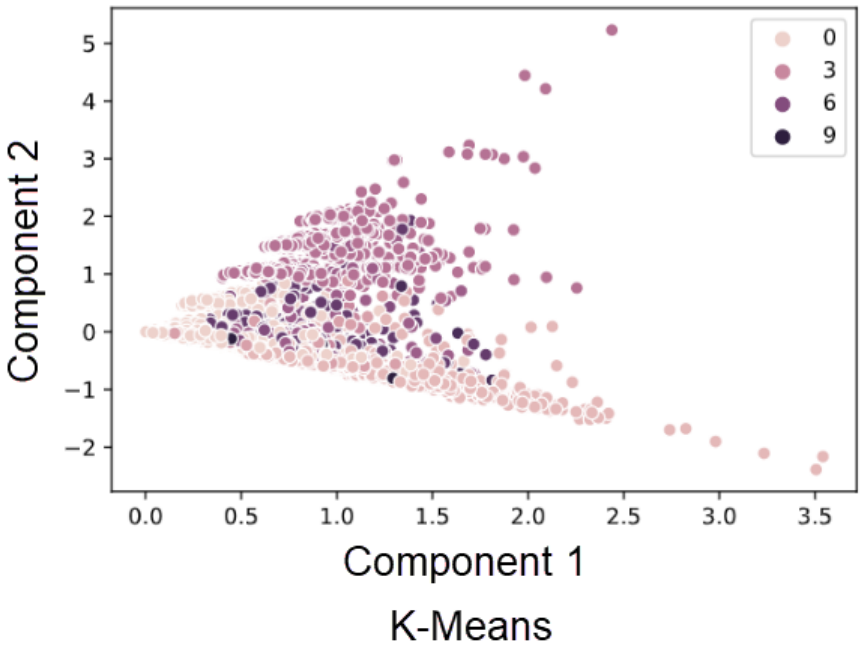


Figure 6.1: Number of Tweets Per Cluster. Both figures show the number of tweets per cluster using ten initial clusters and BOW features for the Pre-COVID dataset.



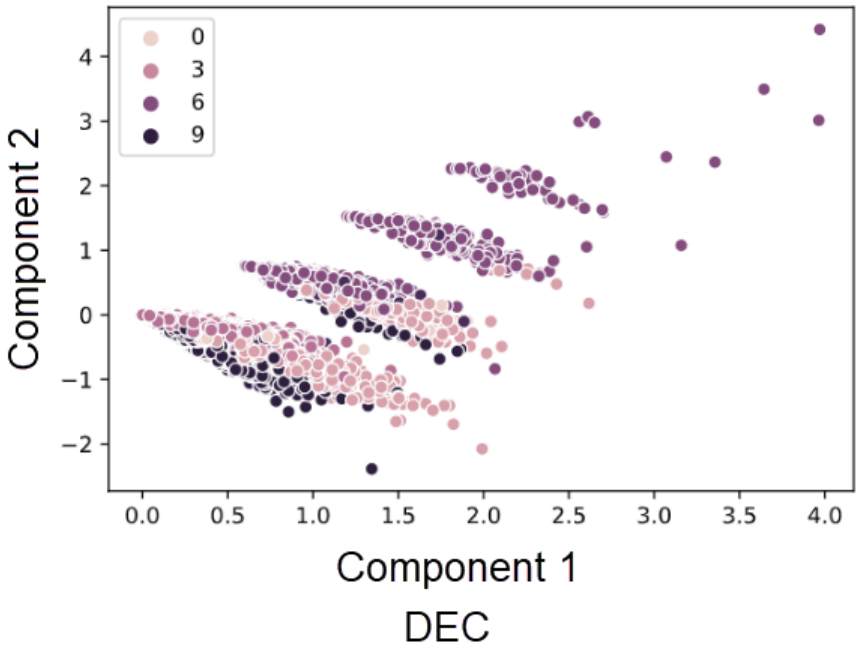


Figure 6.2: Pre-COVID Dataset Cluster Visualization on Reduced Dimensions Using SVD. SVD is used to reduce the clusters (0 to 9) to two dimensions to better visualize the frame groupings.

CHAPTER 7

QUALITATIVE RESULTS

The objective of this chapter is to explore how cryptocurrency frames change over time and their correlation with cryptocurrency day trading behavior. Section 7.1 shows the effects of the pandemic on day trading discussions and behaviors. An analysis between frames before and during the pandemic is also conducted. Section 7.2 discusses how day trading behaviors (e.g., buy, sell, hold) are framed.

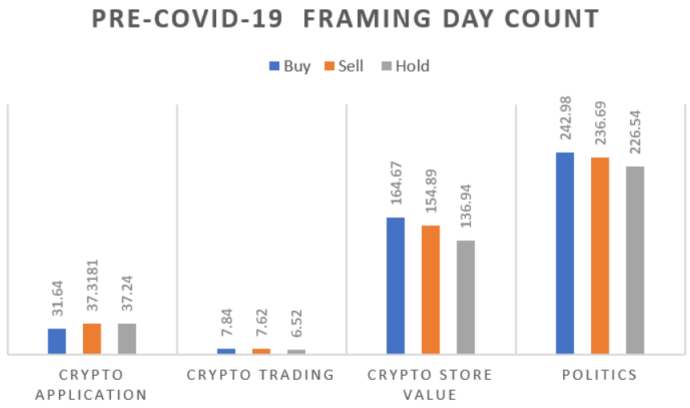
7.1 Frames Before and During the Pandemic

Tables 7.1 and 7.2 show the most frequent words appearing in each of the four clusters extracted from the Pre-COVID or COVID Dataset, respectively. Prior to the pandemic, Table 7.1 shows that the cryptocurrency tweets were framed in terms of aspects important to cryptocurrency itself, i.e., trading actions, applications or uses, and long term store value. Table 7.2 shows that once the pandemic was occurring, the focus of discussion shifted. People still discussed cryptocurrency in terms of trading and applications, however, there was a shift from focusing on long term value and political effects on cryptocurrency to sentiment concerning cryptocurrency and the pandemic. Several tweet examples along with their respective frames during the pandemic are presented in Table 7.3.

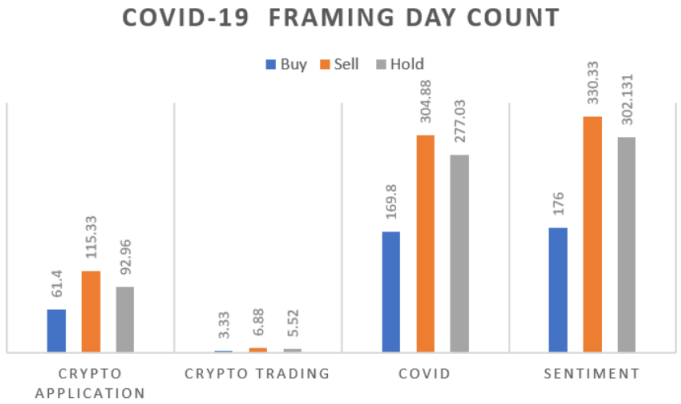
Frame	Most Frequent Words
Crypto Trading	price, bitcoin, usd, market, trading, value, action
Crypto Application	blockchain, btc, business, use, tech, crypto
Crypto Store Value	bitcoin, people, need, want, use, market, value, years
Political	world, man, president, america, china, work, government, time

Table 7.1: Most Frequent Words Per Cluster Prior to COVID-19 (Pre-COVID Dataset).

One interesting event captured by the *Trading* frame in the COVID-19 Dataset was the BTC halving event on May 11, 2020. This halving marks the first quarter of the year as a historical event in the cryptocurrency world because this is the third halving to take place. The past two times



Pre-COVID Frames and Predicted Movement



COVID Frames and Predicted Movement

Figure 7.1: Frames and Movement. Each figure shows the quantity of tweets using a certain frame (separated by a grey line) associated with each investment movement action: buy, sell, or hold.

that halving occurred, Bitcoin later experienced an all-time high price jump. Tracking frames, specifically the *Trading Frame*, and using them to predict a price jump if/when it occurs, or other influential events, is a potential future work that would help guide investor's actions.

7.2 Frames and Momentum Patterns

From observing the frames and momentum patterns prior to the pandemic shown in Figure 7.1, it is notable that *Store Value* frames have a higher frequency when the momentum pattern suggests a *Buy* movement. This correlation makes sense because if there is a belief that some asset will store

Frame	Most Frequent Words
Crypto Trading	money, crypto, btc, trading, finance, investment, halving
Crypto Application	btc, crypto, time, right, know
Sentiment	like, look, things, dont, good, time, feel
Covid	people, coronavirus, covid, pandemic, bitcoin, world, dont

Table 7.2: Most Frequent Words Per Cluster During COVID-19 (COVID Dataset).

Cluster	Tweet Content
Crypto	there is now 2.5x as much BTC on Ethereum as on @Blockstream's Liquid
Crypto	I think he's just using AWS as a useful reference point to explain a cool property of ethereum, rather than suggesting they're substitutes for one another
Crypto	Because it is survivable the remaining miners would have a very strong incentive to stick it out and emerge on the other side 4x more profitable in BTC terms.
Covid	When you digest the sheer size of the 3m+ unemployed who lost jobs this week... Now remember they also lost their healthcare, because it's tied to employment. In the middle of a pandemic.
Covid	Why do you think globalization causes pandemics? This is why I kept asking Preston if he was advocating for ending all international travel.
Covid	"New York City Mayor Bill de Blasio said Monday that New York Police Department officers will pull people out of crowded subway trains as the city continues to grapple with the coronavirus pandemic. "Slippery, slippery slope.
Sentiment	Today is a good day to bring up Betteridge's law of headlines: "Any headline that ends in a question mark can be answered by the word no".
Sentiment	Looks like some things will be made in America again.
Sentiment	Mine too. Buckled up. Ready for launch countdown.

Table 7.3: Example of Tweets Per Cluster Type During the COVID-19 Timeframe.

value it creates more confidence in buying and holding the cryptocurrency.

It is also not surprising that there is an increase in *Political* frames associated with the *Buy* movement. Countries and economies often cited as being politically unstable, such as Botswana, Ghana, Venezuela, and India, have seen an increase in BTC interest because it is more stable than fiat currencies from those countries [1].

[1] https://news.coinsquare.com/government/government-instability-bitcoin/; https://www.un.org/africarenewal/magazine/april-2018-july-2018/africa-could-be-next-frontier-cryptocurrency

Another potential association with the slight increase in *Political* frames during a *Buy* movement is the increase of government adoption and additional regulation of cryptocurrencies. These patterns suggest that prior to the pandemic, if Twitter cryptocurrency discussions were framed in terms of store value or politics, an investor might consider buying more cryptocurrency.

During the COVID-19 time span (Figure 7.1), all frames decrease during an indicated *Buy* movement. However, the opposite occurs, i.e., all frames increase, when the indicated movement is to *Sell*. Regarding both *Trading* and *Application* frames it makes sense to purchase cryptocurrency when nobody is talking about it, and sell it when the interest in those topics rises. The *COVID* frame having a lower frequency during a *Buy* movement could indicate that investors feel less threatened by the market instability introduced by the pandemic, which is the opposite of the general sentiment of investors dealing with physical stock exchange markets.

CHAPTER 8

DISCUSSION

8.1 Conclusion

Predicting day trading behavior, i.e., whether to buy or sell stock, is a complicated task, especially in a volatile asset such as cryptocurrency. This thesis has presented dual modeling pipelines for day trading behavior and framing prediction. The novel results of this thesis demonstrate that language can be used to successfully model cryptocurrency trading behavior and understand how a topic is discussed, or framed.

The first model aims to predict day trading behavior based on daily tweets of influential sources, such as the media or well-known investors. Using classic NLP techniques such as Bag-of-Words and cosine similarity between the DistilBERT representations of tweet features and cryptocurrency tweets, this thesis provides a weakly-supervised model that is capable of distinguishing between day trading actions such as buy, sell, or hold to guide personal investment. Using language-based features, the modeling approach of this work was able to achieve an accuracy of 88.78% with an RNN over a 49.72% Naive Bayes traditional baseline.

The second model focuses on extracting frames to understand how the way influential people and news sources frame cryptocurrency discussions on Twitter affects cryptocurrency day trading. To this end, this thesis has presented an application of an unsupervised deep clustering approach to reveal the latent frames used to discuss day trading behaviors in microblogs. This work is the pioneer in presenting cryptocurrency and trading related frames. Additionally, this thesis presents novel findings which show interesting correlations between investment actions and how cryptocurrency discussions are framed on Twitter, as well as how these framing patterns changed in response to the COVID-19 pandemic.

Across both modeling pipelines, this thesis has contributed: the most accurate machine learning models for studying cryptocurrency discourse on Twitter, the most representative features for day

trading behavior prediction and cryptocurrency framing prediction, and the generation of a new Cryptocurrency Tweets Dataset that contains various features such as daily price movements and content dated prior to and during the COVID-19 pandemic.

8.2 Future Work

This thesis has introduced for the first time in NLP literature the task of cryptocurrency trading framing prediction. This leaves open many avenues for future work to explore. One idea is to improve the framing extraction by continuing to explore the best features for the deep embedded clustering model. Instead of continuing to rely on the machine learning model-extracted frames, more effort could be directed towards creating a larger human annotated corpus of both day trading prediction and cryptocurrency frame databases. Currently, there is no work analysing online discourse and its effect on public opinion for stock market trends, which is also a potential further development of this work that could combine both cryptocurrency and stock market frames.

Cryptocurrencies and blockchain are relatively new concepts that have been gaining popularity rapidly in recent years, and this new technology is revolutionizing and shaping the future of many industries, not just the banking sector. Studying and understanding how people talk about cryptocurrencies, through frames and NLP analysis, is essential to navigating the fast paced changes and impacts introduced by blockchains into societies around the world.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] 2020. The bitcoin volatility index price and more. https://www.bitpremier.com/volatility-index.

[2] 2020. Bvol24h charts and quotes. https://www.tradingview.com/symbols/BVOL24H/.

[3] 2020. Conditional random field. https://en.wikipedia.org/wiki/Conditional_random_field.

[4] 2020. Loss functions¶. https://ml-cheatsheet.readthedocs.io/en/latest/loss_functions.html.

[5] 2020. Understanding lstm networks. https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[6] 2020. Xgboost documentation¶. https://xgboost.readthedocs.io/en/latest/index.html.

[7] Abraham, Jethin, Daniel Higdon, John Nelson & Juan Ibarra. 2018. Cryptocurrency price prediction using tweet volumes and sentiment analysis. In *SMU Data Science Review: Vol. 1: No. 3, Article 1*, .

[8] Abu-Jbara, Amjad, Ben King, Mona Diab & Dragomir Radev. 2013. Identifying opinion subgroups in arabic online discussions. In *Proc. of acl*, .

[9] Amidi, Afshine & Shervine Amidi. 2020. Deep learning cheatsheet star. https://stanford.edu/~shervine/teaching/cs-229/cheatsheet-deep-learning.

[10] Baumer, Eric, Elisha Elovic, Ying Qin, Francesca Polletta & Geri Gay. 2015. Testing and comparing computational approaches for identifying the language of framing in political news. In *In proc. of naacl*, .

[11] Bishop, Christopher M. 2006. *Pattern recognition and machine learning (information science and statistics)*. Berlin, Heidelberg: Springer-Verlag.

[12] Bollen, Johan, Huina Mao & Alberto Pepe. 2011. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proc. of aaai conference on weblogs and social media*, .

[13] Boydstun, Amber, Dallas Card, Justin H. Gross, Philip Resnik & Noah A. Smith. 2014. Tracking the development of media frames within and across policy issues, .

[14] Brownlee, Jason. 2019. A gentle introduction to cross-entropy for machine learning. https://machinelearningmastery.com/cross-entropy-for-machine-learning/.

[15] Brownlee, Jason. 2020. Understand the impact of learning rate on neural network performance. https://machinelearningmastery.com/understand-the-dynamics-of-learning-rate-on-deep-learning-neural-networks/.

[16] Burch, L., E. Frederick & A. Pegoraro. 2015. *Journal of Broadcasting Electronic Media* 59. 399 – 415.

[17] Burch, Lauren M., Evan L. Frederick & Ann Pegoraro. 2015. Kissing in the carnage: An examination of framing on twitter during the vancouver riots. *Journal of Broadcasting & Electronic Media* 59(3). 399–415. doi:10.1080/08838151.2015.1054999. http://dx.doi.org/10.1080/08838151.2015.1054999.

[18] Canellis, David. 2019. Bitcoin has nearly 100,000 nodes, but over 50% run vulnerable code. https://thenextweb.com/hardfork/2019/05/06/bitcoin-100000-nodes-vulnerable-cryptocurrency/.

[19] Card, Dallas, Amber E. Boydstun, Justin H. Gross, Philip Resnik & Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proc. of acl*, .

[20] Chris & Rod Fuentes. 2020. Getting out of loss plateaus by adjusting learning rates. https://www.machinecurve.com/index.php/2020/02/26/getting-out-of-loss-plateaus-by-adjusting-learning-rates/.

[21] CRS, Congressional Research Service. 2020. Global economic effects of covid-19 https://www.who.int/news-room/detail/29-06-2020-covidtimeline.

[22] Derakhshan, Ali & Hamid Beigy. 2019. Sentiment analysis on stock social media for stock price movement prediction. In *Engineering applications of artificial intelligence*, .

[23] Dertat, Arden. 2017. Applied deep learning - part 3: Autoencoders. https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798.

[24] Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

[25] Field, Anjalie, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky & Yulia Tsvetkov. 2018. Framing and agenda-setting in russian news: a computational analysis of intricate political strategies.

[26] Fulgoni, Dean, Jordan Carpenter, Lyle Ungar & Daniel Preotiuc-Pietro. 2016. An empirical exploration of moral foundations theory in partisan news sources. In *Proc. of lrec*, .

[27] Fumo, David. 2017. A gentle introduction to neural networks series - part 1. https://towardsdatascience.com/a-gentle-introduction-to-neural-networks-serie-part-1-2b90b87795bc.

[28] Gao, B., Y. Yang, H. Gouk & T. M. Hospedales. 2020. Deep clusteringwith concrete k-means. In *Icassp 2020 - 2020 ieee international conference on acoustics, speech and signal processing (icassp)*, 4252–4256.

[29] Goodfellow, Ian, Yoshua Bengio & Aaron Courville. 2016. *Deep learning*. MIT Press. http://www.deeplearningbook.org.

[30] Hadifar, Amir, Lucas Sterckx, Thomas Demeester & Chris Develder. 2019. A self-training approach for short text clustering. In *Proceedings of the 4th workshop on representation learning for nlp (repl4nlp-2019)*, 194–199. Florence, Italy: Association for Computational Linguistics. doi:10.18653/v1/W19-4322. https://www.aclweb.org/anthology/W19-4322.

[31] Harlow, Summer & Thomas Johnson. 2011. The arab spring| overthrowing the protest paradigm? how the new york times, global voices and twitter covered the egyptian revolution. *International Journal of Communication* 5(0).

[32] Hasan, Kazi Saidul & Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proc. of emnlp*, .

[33] Hastle T., Friedman J., Tibshirani R. 2009. *The elements of statistical learning*. Springer-Verlag New York.

[34] Hinton, Geoffrey E., Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever & Ruslan Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR* abs/1207.0580. http://arxiv.org/abs/1207.0580.

[35] Hochreiter, Sepp & Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9(8). 1735–1780. doi:10.1162/neco.1997.9.8.1735.

[36] Jain, Arti, Shashank Tripathi, Harsh Dhardwivedi & Pranav Saxena. 2018. Forecasting price of cryptocurrencies using tweets sentiment analysis, .

[37] Jang, S. Mo & P. Sol Hart. 2015. Polarized frames on "climate change" and "global warming" across countries and states: Evidence from twitter big data. *Global Environmental Change* 32. 11–17.

[38] Jelodar, Hamed, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li & Liang Zhao. 2019. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications* 78(11). 15169–15211.

[39] Johnson, Kristen, Di Jin & Dan Goldwasser. 2017. Leveraging behavioral and social information for weakly supervised collective classification of political discourse on twitter. In *Proc. of acl*, .

[40] Jurafsky, Daniel & James H. Martin. 2020. *Speech and language processing(3rd ed. draft)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

[41] Kendall, M. G., A. Stuart & J. K. Ord. 1987. *Kendall's advanced theory of statistics*. USA: Oxford University Press, Inc.

[42] Klemens, Sam. 2020. How many bitcoins are left? (updated 2020). https://www.exodus.io/blog/how-many-bitcoins-are-left.

[43] Kouloumpis, Efthymios, Theresa Wilson & Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! In *Proc. of aaai conference on weblogs and social media*, .

[44] Kuepper, Justin. 2020. Volatility. https://www.investopedia.com/terms/v/volatility.asp.

[45] Kyröläinen, Petri. 2008. Day trading and stock price volatility. *Journal of Economics and Finance* 32. 75–89. doi:10.1007/s12197-007-9006-2.

[46] Lafferty, John D., Andrew McCallum & Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning* ICML '01, 282–289. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

[47] Lewis, David D. 1998. Naive (bayes) at forty: The independence assumption in information retrieval. In Claire Nédellec & Céline Rouveirol (eds.), *Machine learning: Ecml-98*, 4–15. Berlin, Heidelberg: Springer Berlin Heidelberg.

[48] Li, Tianyu Ray, Anup S. Chamrajnagar, Xander R. Fong, Nicholas R. Rizik & Feng Fu. 2019. Sentiment-based prediction of alternative cryptocurrency price fluctuations using gradient boosting tree model. *Frontiers in Physics* 7. 98.

[49] Meraz, Sharon & Zizi Papacharissi. 2013. Networked gatekeeping and networked framing on #egypt. *The International Journal of Press/Politics* 18(2). 138–166.

[50] Mitchell, Thomas M. 1997. *Machine learning*. USA: McGraw-Hill, Inc. 1st edn.

[51] Mone, Lesa. 2019. I read crypto twitter for hours daily — here are the 40 accounts that really matter. In *Consensys blog*, https://bit.ly/36I0tiC.

[52] Nakamoto, Satoshi. 2008. Bitcoin: A peer-to-peer electronic cash system. www.bitcoin.org.

[53] Oludare Isaac Abiodun, Abiodun Esther Omolara Kemi Victoria Dada Nachaat AbdElatif Mohamed Humaira Arshadf, Aman Jantan. 2018. State-of-the-art in artificial neural network applications: A survey. *Heliyon* 4.

[54] Ortega, Joaquín, Nelva Almanza-Ortega, Andrea Vega-Villalobos, Rodolfo Pazos-Rangel, José Crispin Zavala-Diaz & Alicia Martínez-Rebollar. 2019. *The k-means algorithm evolution*. doi:10.5772/intechopen.85447.

[55] Partz, Helen. 2019. Number of americans owning crypto doubled in 2019: Finder. In *Coin telegraph*, https://cointelegraph.com/news/number-of-americans-owning-crypto-doubled-in-2019-finder.

[56] Rajaraman, Anand & Jeffrey David Ullman. 2011. *Data mining* 1–17. Cambridge University Press. doi:10.1017/CBO9781139058452.002.

[57] Rao, Tushar & Saket Srivastava. 2012. Analyzing stock market movements using twitter sentiment analysis. In *Proc. of international conference on advances in social networks analysis and mining*, .

[58] Reiff, Nathan. 2020. Blockchain explained. https://www.investopedia.com/terms/b/blockchain.asp.

[59] Reiff, Nathan. 2020. Why bitcoin has a volatile value. https://www.investopedia.com/articles/investing/052014/why-bitcoins-value-so-volatile.asp.

[60] Ritter, Alan, Colin Cherry & Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Proc. of naacl*, .

[61] Sanh, Victor, Lysandre Debut, Julien Chaumond & Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

[62] Sanh, Victor, Lysandre Debut, Julien Chaumond & Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR* abs/1910.01108. http://arxiv.org/abs/1910.01108.

[63] Si, Jianfeng, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li & Xiaotie Deng. 2013. Exploiting topic based twitter sentiment for stock prediction. In *Proc. of 51st annual meeting of the association for computational linguistics*, .

[64] Sridhar, Dhanya, James Foulds, Bert Huang, Lise Getoor & Marilyn Walker. 2015. Joint models of disagreement and stance in online debate. In *Proc. of acl*, .

[65] Tsur, Oren, Dan Calacci & David Lazer. 2015. A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *Proc. of acl*, .

[66] Vidal, Tiago. 2020. How traders can use twitter to anticipate bitcoin price moves, volume, .

[67] Walczak, Steven. 2001. An empirical analysis of data requirements for financial forecasting with neural networks. *Journal of management information systems* 17(4). 203–222.

[68] Walker, Marilyn A., Pranav Anand, Robert Abbott & Ricky Grant. 2012. Stance classification using dialogic properties of persuasion. In *Proc. of naacl*, .

[69] West, Robert, Hristo S Paskov, Jure Leskovec & Christopher Potts. 2014. Exploiting social network structure for person-to-person sentiment analysis. *TACL* .

[70] Xie, Junyuan, Ross Girshick & Ali Farhadi. 2015. Unsupervised deep embedding for clustering analysis .

[71] Xie, Junyuan, Ross Girshick & Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis, vol. 48 Proceedings of Machine Learning Research, 478–487. New York, New York, USA: PMLR. http://proceedings.mlr.press/v48/xieb16.html.