

Analysis of Cryptocurrency Public Sentiment Shifts

MSc Research Project
FinTech

Yen Lyn Ooi
Student ID: X19128657

School of Computing
National College of Ireland

Supervisor: Victor Del Rosal

National College of Ireland
MSc Project Submission Sheet
School of Computing



Student Name: Yen Lyn Ooi

 X19128657

Student ID:
 MSc FinTech 2020

Programme: **Year:**
 Research Project

Module:
 Victor Del Rosal

Supervisor:
Submission Due Date: 17/8/2020

Project Title:
 Analysis of Cryptocurrency Public Sentiment Shifts

Word Count: 6664 **Page Count:** 21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:

Date:

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	<input type="checkbox"/>
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	<input type="checkbox"/>
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	<input type="checkbox"/>

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Analysis of Cryptocurrency Public Sentiment Shifts

Yen Lyn Ooi
X19128657

Abstract

The popularity and emergence of digital currency can be attributed to social media as it has a large user base for online discussion. The crypto market is highly dependent on socially constructed opinions as investors rely on online sources to acquire related information. A headline is considered as a summary of an article in a single sentence as it is the first that can gain a reader's attention. The objective of this paper is to analyze the trend of cryptocurrency headlines in terms of positive and negative sentiment. The dataset is obtained from the Kaggle webpage containing news headlines from five online platforms. The paper will be using a lexicon-based sentiment approach to identify the binary sentiment. Support Vector Machine algorithm will be used to evaluate and optimize the sentiment results. This paper contributes to the gap in the literature by providing an empirical analysis of overview changes of the cryptocurrencies with multiple news platforms and longer periods. The findings of the paper resulting in negative sentiment for the trend of headlines, but it showed a different view in terms of positive polarity. Furthermore, the TD-IDF model outperformed the sentiment model in SVM modeling.

Keywords: Cryptocurrency, Sentiment analysis, Lexicon sentiment, Natural Language Processing (NLP), Support Vector Machine (SVM).

1 Introduction

During the 90s tech boom, there have been many attempts at creating digital currency such as Flooz, Beenz, and DigiCash but failed due to security and centralization issues¹. Thereafter, the development of the electronic cash system was a lost cause for a period. In 2009, anonymous Satoshi Nakamoto first introduced Bitcoin in the proposed white paper, a decentralized peer-to-peer electronic system without relying on central authority (Nakamoto, 2009). Cryptocurrencies are so-called because their consensus-keeping process is safeguarded with strong cryptography. Every transaction consists of public keys that buyers and sellers exercised to transfer coins publicly while private keys are known only to the user. According

¹ <https://www.coindesk.com/3-pre-bitcoin-virtual-currencies-bit-dust>

to Coin ATM Radar, there are currently 8,539 Crypto ATMs in 73 countries². As Bitcoin is becoming more widespread acceptance, cryptocurrencies can be used as payment for goods and services as well as an investment option that is similar to gold, cash, or gift cards. Since the first introduction of Bitcoin, many more alternative coins also known as altcoins have been launched in the crypto market. The altcoins shared the underlying blockchain technology but were employed with different algorithm designs. The recent coronavirus pandemic has pushed the process of cryptocurrency into many more people's surroundings than ever before. The recent news reported that China has been testing on the national digital currency in the cities of Shenzhen, Chengdu, Suzhou, and Xiongan³. The central banks of England, Japan, Sweden, Switzerland, and the EU are also working on central bank digital currencies (CBDC)⁴.



Figure 1: The total number of cryptocurrencies ATMs worldwide.

1.1 Research Background & Motivation

The popularity and emergence of digital currency can be attributed to social media as it has a large user base for online discussion. Digital currencies are different from equities, there is no consensus on how to value a cryptocurrency, thus the topic discussions are widely varied. Because of that reason, the crypto market is highly reliant on socially constructed opinions. The average crypto investors are young and inexperienced, they tend to rely on online resources to obtain information about the features of coins⁵. Individuals are able to gain partial ownership of that company when investing in the stock market, no matter how small the stake is. In addition to that, investors able to review and analyze companies' balance sheets and earnings reports to determine an appropriate value. In contrast to traditional market, digital currencies trading has no government backing, there is practically nothing tangible to base on the price movements. Thus, most of the cryptocurrency's investors are speculators, they forecast the movement of price based on the historical charts.

² <https://coinatmradar.com/>

³ <https://news.bitcoin.com/digital-yuan-to-fuel-chinas-economic-reign-mcdonalds-starbucks-subway-test-pbocs-cryptocurrency/>

⁴ <https://www.independent.co.uk/life-style/gadgets-and-tech/news/cryptocurrency-bank-of-england-digital-currency-a9295136.html>

⁵ <https://bitcoinmagazine.com/articles/new-research-shows-how-young-adults-drive-the-cryptocurrency-revolution>

The extensive literature reviews revealed the impact of social media platforms that could influence the cryptocurrencies. Lamon *et al.*, (2017) forecasted price fluctuations of cryptocurrencies from news and social media tweets online by using different machine learning approaches such as Logistic Regression, Naïve Bayes, and Linear Support Vector Machine (SVM). Logistics Regression presented the best performance along with relevance text data and tweets. Dulău and Dulău (2019) analyzed the relevance of cryptocurrencies text posted on Reddit and Twitter platforms by using sentiment analysis. In a recent study, Narman and Uulu (2020) examined users' comments in terms of positive and negative of six cryptocurrencies on Reddit. One of the notable results showed that the movement of coins prices would change along with the number of comments. Furthermore, there is currently a scarcity of studies that have explored multiple news sources in the crypto market. Therefore, this paper contributes to the gap in the literature by extending the analysis of cryptocurrency to multiple platforms with longer periods. It would provide an overview of changes on the topic of cryptocurrency. Investors and consumers will get to understand more about it and thus could be a valuable guide to investors and the general public for their investment option.

The objective of this paper is to analyze cryptocurrencies headlines in terms of positive and negative as well as identify the most occurring words from online news publications. The dataset that we have obtained is from the Kaggle webpage covering five years of data, from 2013 to 2018. The paper is accomplishing through quantitative research with a lexicon-based dictionary to identify sentiment from the news headlines along with a machine learning algorithm for optimization.

The significant effect of online platforms motivated for the research question:

“How has sentiment towards cryptocurrencies shifted over the years as evidenced in public media and news?”

The contents of the paper are structured as follows. Section 2 reviews the related works. Section 3 presents the methodology. Section 4 demonstrates the design specification. Section 5 describes how the approach was implemented. Section 6 discusses the findings and results. Last, section 7 concludes the paper with future works.

2 Related Work

2.1 Cryptocurrencies in Social Media

There are several past studies regarding the analysis of cryptocurrency with different methods from online news data, and the focus was exclusively on predicting prices. Gurriba *et al.*, (2019) investigated the impact of return towards cryptocurrencies in the major news announcement by applying the approach of Granger causality and impulse responses. The study focused on the United States (US), United Kingdom (UK), and the European country which has the highest number of wallet users and providers. Wooley *et al.*, (2019) predicted the cryptocurrency prices on the Reddit community. Kim *et al.*, (2017) extracted keywords from the Bitcoin online forum to forecast the price of the currency. Dulău and Dulău (2019) applied

sentiment analysis from Twitter and Reddit by using two dictionary-based approaches: Stanford CoreNLP and IBM Watson in the research area. Glenski *et al.*, (2019) highlighted the interest spread of Ethereum is smaller than Bitcoin and Monero, whereas the discussion of Monero is popular than Bitcoin on Reddit.

Kim *et al.*, (2016) predicted cryptocurrencies prices and the number of transactions extracting from user comments in the period of 2013 to 2016. Ripple and Ethereum proved to be associated with negative comments online whereas positive comments significantly influence the price movements of Bitcoin. Besides, the number of transactions determined to be significantly associated with user replies rather than users' posted comments. By studying the Twitter interaction of cryptocurrencies through social network analysis, Park and Lee (2019) revealed the social interaction surrounding cryptocurrencies can provide a better understanding of the dynamic and complexity of the cryptocurrency market as well as allowing a more accurate valuation. It is worth noting in the findings that nearly all the negative tweets' messages disappeared when the price of most cryptocurrency shoots up in 2017 while positives messages with encouragement and support also increased rapidly during that year. In the following year, the positive tweets reduced after the price of the cryptocurrency has reached its peak.

Valencia *et al.*, (2019) extended time series analysis along with machine learning and sentiment analysis using Twitter data. By applying different approaches including Neural Networks, SVM, and Random Forests, Litecoin is found to be the most predictable currency in the market with the highest precision score, followed by Bitcoin and Ripple. Steinert and Herff (2018) focused on analyzing the return of altcoins using social media data. The paper attempted to explore the data from different perspectives including the extent of Twitter activity quantified, the number of tweets, and the sentiment score. Kang *et al.*, (2019) analyzed the topic of cryptocurrencies in the Bitcoin forum and found that most users are interested in price information whereas social media users are not only interested in the price but also in other information. Another interesting finding is that the existence of opinion leaders in a forum facilitates processing information related to Bitcoin. Rognone *et al.*, (2020) examined the reaction of news sentiment between Bitcoin and six traditional currencies. The findings showed that Bitcoin reacts positively to both good and bad news, suggesting investor enthusiasm for Bitcoin regardless of the news sentiment. According to the papers reviewed, we can conclude that the usage of Twitter, the Bitcoin forum, and Reddit data influenced the fluctuations of prices in the market. Besides, the prediction accuracy of Ripple and Litecoin is consistently low due to the difficulties of acquiring relevant data during the past periods.

2.2 Crypto Market

Gidea *et al.*, (2020) combined topological data analysis (TBA) with a machine learning technique to analyze the behavior of four major cryptocurrencies namely Bitcoin, Ethereum, Litecoin, and Ripple before the 2018 digital asset market crash. By applying TBA to the time series of each cryptocurrency, it successfully identified early warning signals of the market crash that are consistent with the observed data. Thus, suggesting the approach can recognize the changes in the relevant time series before the crash. In exploring the rapid growth of the

social and news media landscape, Gan *et al.*, (2020) highlighted that social media is becoming the dominant media source in the financial market. Donmez *et al.*, (2020) that attempts to investigate the relations between cryptocurrencies by using a full historical price. The findings showed that Bitcoin remains an important central role in the cryptocurrency market. BTC and LKY occupied the biggest cluster in the hierarchical tree indicated a highly correlated relationship for both digital currencies.

In the early studies, Ciaian *et al.*, (2016) analyzed the characteristics of Bitcoin that may become a global currency. It was shown that price volatility is one of the largest differences which differ substantially from traditional currencies thus constrain its expansion globally. It generates uncertainty for investors due to its inability to maintain a stable rate over time and may fail to accurately convey the relative price of goods and services in the economy. There is another corresponding paper David *et al.*, (2017) that also attempted to explore the possibility of cryptocurrencies being a new investment opportunity in the financial market. It is concluded that cryptocurrency can be a good option to help diversify risk in a traditional portfolio due to the correlations between the cryptocurrencies. Traditional assets are consistently low and yet cryptocurrencies have higher risk hence higher average return to the investor. By applying the GARCH modeling, Chu *et al.*, (2017) suggested cryptocurrencies such as Bitcoin, Ethereum, Litecoin, and many others are high volatility which is suitable for risk-seeking investors. The findings are similar to Sajter (2019) which focusing on Bitcoin, Ethereum, and Ripple using time series analysis.

2.3 Lexicon Building

Lexicon sentiment is one of the techniques used to analyze the text document in sentiment analysis. Social media and news articles are the most reliable way of extracting valuable information through a bunch of text compared to other platforms. Im *et al.*, (2015) focused on rule-based sentiment analysis to investigate the financial news articles with positive and negative sentiment. Asghar *et al.*, (2017) applied SentiWordNet based classification to classify the reviews in the online forum. Taboada *et al.*, (2011) used another technique, Semantic Orientation CALculator (SO-CAL) to perform polarity classification from the texts. Dhaoui *et al.*, (2017) conducted comparative research between the lexicon-based approach and machine learning approach on social media comments. Linguistic Inquiry and Word Count (LIWC) are applied in lexicon-based methods to study the sentiment across different domains and are also used supervised machine learning to train the classifiers and perform the classification sentiment. The results showed that both methods achieve similar values of F-score in terms of negative and positive classification. The paper further demonstrated a combined approach of both lexicon and machine learning which indicated a higher accuracy in positive sentiment while the value in negative sentiment remains the same. The findings suggesting that the combined approach is particularly useful for marketers in identifying positive text accurately.

Augustyniak *et al.*, (2016) employed dictionary-based lexicons for sentiment classifier and decision tree machine learning as a fusion classifier based on the lexicon's output. The proposed method reduced the computation time needed for the computer system whereas the overall accuracy remains the same. However, there is an argument regarding the existence of bias in

sentiment that led to poor sentiment score performance in the analysis. Han *et al.*, (2018) introduced the determination of weight and threshold parameters to properly weighted the polarity of the Amazon product review by balancing the number of positive and negative reviews in the training data set to get an optimal result. Overall, the strategy showed improvement in the performance of the lexicon sentiment with better accuracy and F-measure in the experiments. There are several papers (Zainuddin *et al.*, 2017; Kaur and Kumari, 2016; Ersahin *et al.*, 2019) presented a hybrid method using lexicon approach and supervised learning methods in sentiment analysis and all showed improved accuracy result in future prediction.

Table 1: Summary of articles in Lexicon building

Authors	Year	Data collection	Lexicon Approach
Taboada <i>et al.</i> ,	2011	Mechanical Turk	Semantic Orientation CALculator (SO-CAL)
Im <i>et al.</i> ,	2015	Financial news articles	Stanford RNN parser
Augustyniak <i>et al.</i> ,	2016	Amazon reviews	Dictionary-based & Machine learning algorithms
Asghar <i>et al.</i> ,	2017	Forum reviews	SentiWordNet
Dhaoui <i>et al.</i> ,	2017	Social media comments	Linguistic Inquiry & Word Count
Han <i>et al.</i> ,	2018	Amazon product review	SentiWordNet
Ersahin <i>et al.</i> ,	2019	Movie, Hotel and Tweet reviews	SentiTurkNet & SVM & Naïve Bayes

3 Research Methodology

The methodology followed the Knowledge Discovery in Database (KDD) process created by Piatesky-Shapiro at the workshop in 1989 (Piatesky-Shapiro, 1991). In this paper, the KDD framework is followed due to its stage-wise approach that uncovers the hidden patterns from a large unstructured dataset. Additionally, KDD focuses on the overall process of knowledge to discover from the data that can be interpreted as useful insights. Moreover, this paper is dedicated to text data which will be mining through a text document known as text mining.

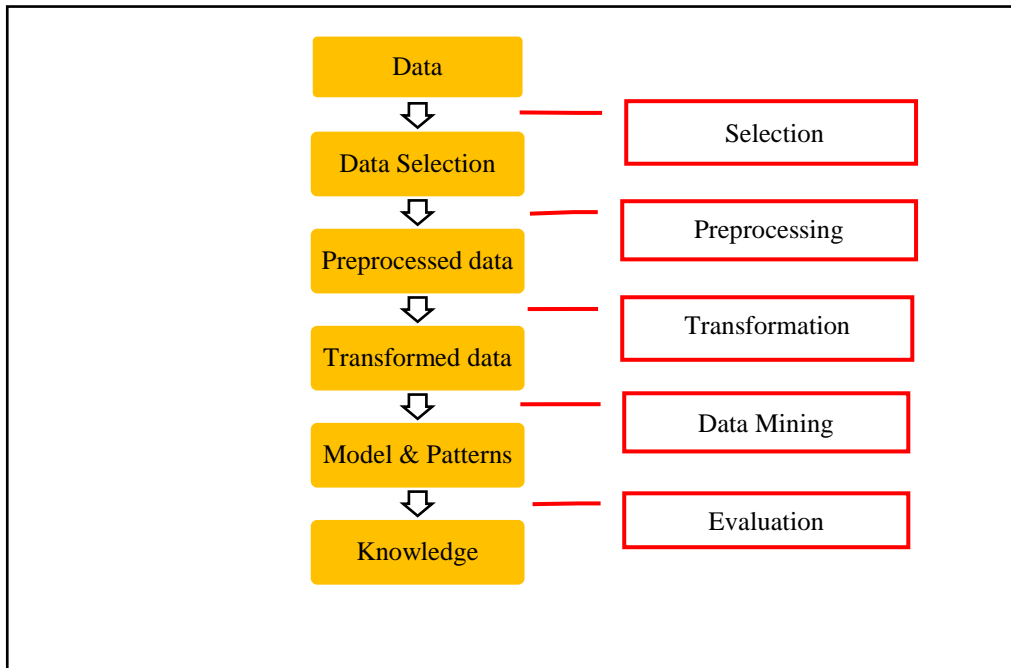


Figure 2: KDD process.

3.1 Data Understanding

This is the preliminary step to understand the problem that attempting to solve. This could be done through researching and gathering prior knowledge, understand the challenges as well as the limitations dealt with the topic. It is also important to concern the goals that are achievable within the time frame.

3.2 Data Selection

With the defined goals, data that aligned with knowledge discover should be determined. This includes looking up relevant data from the multiple databases and gathers samples for validation. Thereafter, combining all data into one dataset so that significant data are retrieved from the database. This process is significant because text mining learns and exposes insights from the available data. If some important attributes are missing, may lead to failure in the entire study.

3.3 Pre-processing

In this step, data reliability will be enhanced. Raw data are often noise and untidy, it is required to perform data cleaning including dealing with missing values, removal of duplicate values, and outliers. In our case, it should be noted that raw text cannot be analyzed quantitatively which the data must have some numeric representation. In Natural Language Processing, it performs the task on the raw text corpus into a tidy form so that data is appropriate for mining.

3.4 Data transformation

Readable data is now prepared and developed for data mining. This step includes dimensionality transformation such as feature selection and splitting into train and test data. In text mining, we may need to consider the format of vectors and sequences index tables.

3.5 Data Mining

This stage is to decide on the appropriate data mining technique such as regression, classification, and so on. The selected various specific modeling techniques are used to discover and extract patterns from datasets. Prediction and Description are the two major objectives of data mining. Prediction is referred to as supervised learning whereas description includes unsupervised and visualization aspects of learning. There is also a need to employ the algorithm several times to obtain a satisfactory result.

3.6 Evaluation

This segment is to evaluate and interpret the mined result such as the probability of accuracy values concerning the defined objective in the first step. It focuses on the comprehensive of the persuaded model.

3.7 Knowledge

It is now able to incorporate the mined knowledge for further action. The knowledge becomes active in the sense that we may make changes to the system and measures the effects. Through the result evaluation, the study able to gain an interesting insight into users as well as provide research contributions.

4 Design Specification

Natural Language Processing (NLP) deals with the interaction between computer and text languages which can be counted as a domain in Machine Learning. Besides, sentimental analysis is contextual mining of text to extract subjective information from a bunch of words to understand the opinion. Sentiment lexicon is also known as a rule-based approach in which the words are classified into a positive or negative value. It is based on an algorithm with a clearly defined description of an opinion to identify the subjective, polarity score, or the opinion. Lexicon-based approach used the prior polarity lexicon to determine the semantic orientation of the text document. Each of the words in the analyzed text is matched with the words in the subjectivity lexicon. The polarity of the word will be tagged according to its matched pair's polarity if the pair is matched in the lexicon dictionary.

There are mainly three types of lexicon that could be used in sentiment classification. The Bing Liu Lexicon contains a binary categorization model that sort words into negative and positive sentiment (Liu, *et al.*, 2005). In Bing Liu lexicon, there are 2,006 of positive words, and 4,783 of negative words contained in the list of English words. The NRC lexicon from Mohammad and Turney (2013) contains a list of words along with eight basic emotions and two binary

sentiments (positive; negative; anger; fear; anticipation; trust; surprise; sadness; joy; disgust). The AFINN lexicon from Nielsen (2011) contains a list of words ranged with an integer between -5 (negative) and +5 (positive) with 2,477 coded words. These are unsupervised classifiers that can be used for prediction with the following rules. If the text associates with only positive emoticon and it is labeled as positive. In contrast, if the text contains negative emotions, it is labeled as negative. If there are no emoticons in the text, it is classified as neutral. In sentiment analysis, it involves several operations with the text document such as stemming, tokenization, part of speech tagging, parsing, and lexicon analysis.

5 Implementation

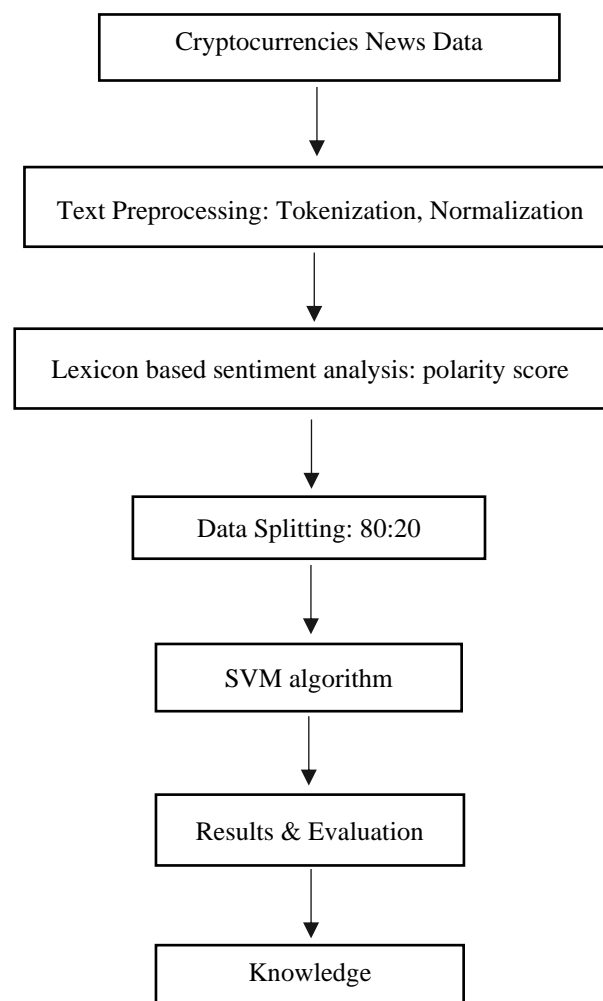


Figure 3: KDD flow.

5.1 Data

The datasets are obtained from the Kaggle webpage to conduct the research. The dataset consists of cryptocurrency information collected from five lists of online media platforms: Cointelegraph, News BTC, CoinDesk, CCN, and Forklog. We chose to analyze two cryptocurrency datasets from 2013 to 2017. Two files are merged into a data frame to have a complete five years of data for easily accessible. Particularly, there are totals of 39,467 instances and 7 attributes included in the data. The seven attributes consist of the URL of five different media news, the title of cryptocurrency news, the contents of the news, hypertext markup language (HTML), year, news author, and sources of news.

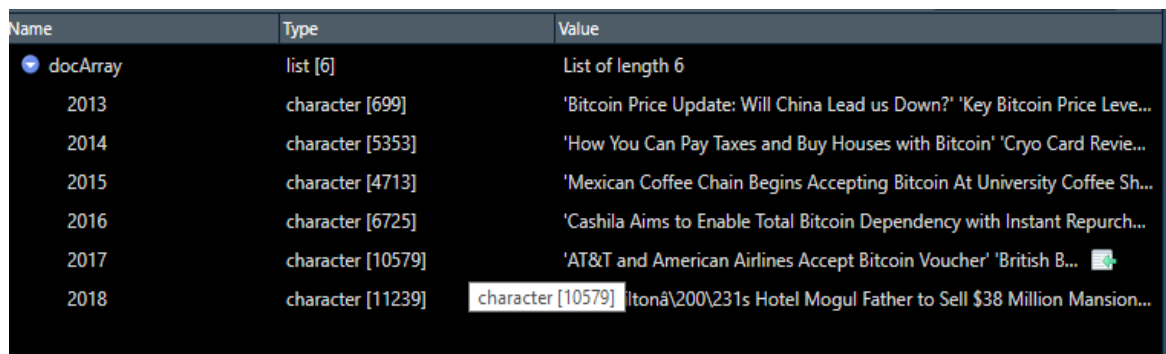
5.2 Pre-processing

5.2.1 Feature Selection

In the crypto dataset, some of the attributes such as title and year are retained and used for our analysis. The purpose of feature selection is to reduce the number of variables in the data by selecting only the most compatible value to help in the analysis process. After conducting the feature selection, the dataset focused on two attributes.

5.2.2 Tokenization

This is a step where the headlines of news of text are broken down into individual words or phrases in a text file. For example, larger chunks of the word can be tokenized into sentences, sentences can be tokenized into words. We created a corpus contained in the *tm* package to explore a stream of words in a sentence.




Name	Type	Value
docArray	list [6]	List of length 6
2013	character [699]	'Bitcoin Price Update: Will China Lead us Down?' 'Key Bitcoin Price Leve...
2014	character [5353]	'How You Can Pay Taxes and Buy Houses with Bitcoin' 'Cryo Card Revie...
2015	character [4713]	'Mexican Coffee Chain Begins Accepting Bitcoin At University Coffee Sh...
2016	character [6725]	'Cashila Aims to Enable Total Bitcoin Dependency with Instant Repurch...
2017	character [10579]	'AT&T and American Airlines Accept Bitcoin Voucher' 'British B... 
2018	character [11239]	character [10579] ltonã\200\231s Hotel Mogul Father to Sell \$38 Million Mansion...

Figure 4: Lists of words.

5.2.3 Normalization

The text needs to be normalized before further processing. There are several steps to perform the corpus by applying the *tm* function under each component. These includes:

- Set the entire news title to lowercase to avoid having the same words since R is case sensitive.
- Remove punctuations and numbers in the text.

- Remove commonly used English stop words such as ‘for’, ‘a’, ‘the’.
- Remove custom stop words “will” as it occupied a large portion in the text documents.
- Remove extra whitespaces.
- Remove special characters such as ‘@%/>’.
- Perform text stemming to obtain its word stem such as 'bitcoins' to 'bitcoin' which helps to increase accuracy in our text and decrease the vocabulary space.

5.3 Sentiment Analysis

A document term matrix is a sparse matrix where each row of the matrix is a document vector, with one column of every term in the entire corpus. It contains the frequency of words that each cell in that matrix will be an integer of the number of times that term was found in the document. Figure 5 shows a better understanding of how the text is formatted with the *inspect()* function. The dataset contains 6 documents with 13,212 terms.

```
> inspect(dtm)
<<TermDocumentMatrix (terms: 13212, documents: 6)>>
Non-/sparse entries: 29418/49854
Sparsity           : 63%
Maximal term length: 19
weighting          : term frequency (tf)
sample            :
  Docs
Terms  1    2    3    4    5    6
analysi 0   34  806  718 1412 1139
bank    39  226  214  441  636  608
bitc    585 3672 2847 3229 4592 3192
ckchain 2    96  348 1615 1699 1611
currenc 32  288  231  317  785 1693
ethereum 0   27   76  775 1115  778
exchang 51  273  188  305  564  822
new     37  296  225  264  614  535
price   51  371 1448 1591 2494 1805
technic 1   10  444  316  889  545
```

Figure 5: Document term frequency.

Tidy text format is created to break the titles of cryptocurrency into individual words. Lexicon-based is unsupervised techniques used to classify the headlines into a binary field of positive and negative. Bing Liu approach is applied in the paper to find out the positive sentiment or negative sentiment of the word. With the automation of counting, the sentiment will be classified as positive when the resulting number is more than zero. In contrast, if the resulting number is less than zero then it will be classified as negative. We determine the overall polarity of the text document in a more precise aspect level. The determination of the polarity of words will be assigned to every entity in the words if the words in the headlines are matched with the Bing’s dictionary. The score will be +1 if it is a positive word, otherwise will be -1 for negative words. The formulas of polarity sentiment as given below.

$$\text{SentiScore} = \text{Positive} - \text{Negative} / (\text{Positive} + \text{Negative}) \quad (1)$$

$$\text{Positive Score} = \text{Positive} / \text{Sum of Sentiment} \quad (2)$$

Furthermore, there are several weighting methods to evaluate the importance of each feature by assigning a certain weight in the text document such as Term Frequency - Inverse Document Frequency (TF-IDF). TF-IDF based features are applied to the Bing lexicon to increase the performance of the model. TF-IDF aims to determine the words that are important in the crypto headline by assigning a lower weight for commonly used words such as 'is' and 'the', and higher weight for words that are least important in a document. The value of TF-IDF is zero for those extremely common words in the text. We smoothen the value by adding one to the equation as we cannot be divided by zero. The equation of TF-IDF is as below.

$$\text{tf-idf}(t) = \text{tf}(t, d) * \log(N/(\text{df}+1)) \quad (3)$$

Where t is the word, d is the document, N is the number of documents.

5.4 Machine Learning Classifier

Support Vector Machine (SVM) was considered for model training due to its great sparseness of handling text data. It is known to be a good classifier for two classes and less computationally demands in categorize new observations (Yu and Nwet, 2020). The dataset with sentiment label and features is divided into a ratio of 80:20 for the accuracy of the sentiment model. The algorithm is used to train a sentiment classifier that is taken based on the frequency of occurrence of various words contained in the document. The purpose of SVM is to find the optimal hyperplane which has a maximum margin. Margin refers to the distance between the hyperplane (lines) with the nearest point from each class which is usually called Support Vector. A linear function is used as kernel function $K(x_i, x_i^1)$ in this paper to handle non-linear problems. The equation for a hyperplane can be seen below.

$$f(x) = wx + b = 0 \quad (4)$$

Where, w represents the vector and b represents the intercept. We can define the margins by changing the b and rescaling the u , as given below.

$$\text{For positive plane: } f(x) = wx + b = +1 \quad (5)$$

$$\text{For negative plane: } f(x) = wx + b = -1 \quad (6)$$

Where, w represents the weight vector, b is the bias, and x is the feature vector. Therefore, SVM minimizes the problem to:

$$\text{minimize } \|w\| \text{ subject to } y_i(wx_i + b) \geq 1 \text{ for } i = 1 \dots N \quad (7)$$

5.5 Evaluation

After applying the KDD process accordingly by utilizing the sentiment analysis and compatible machine learning, we evaluate the results obtained from SVM. The obtained classified sentiment will be presented in various charts to indicate a visual effect and understanding of the data. To measure the performance of SVM, confusion matrix is conducted and focused on the accuracy, sensitivity, and F-measures. The matrix is calculated according to the value of True Positive (TP), True Negative (NP), False Positive (FP), and False Negative (FN). The formulas of the equations are shown below.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (9)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (10)$$

6 Results and Evaluations

6.1 Data visualization

We used the corpus created to build a term-document matrix containing the frequency of the words. Figure 6 presents the most frequently occurred words in the cryptocurrency headlines. It is apparent that Bitcoin is the most secure word in the text. Other than Bitcoin, Ethereum also shown to be one of the popular digital currency in the crypto market as it is under the list of top 5 words. In addition to that, Litecoin and Ripple also appeared to be under the top 40 words in the diagram.

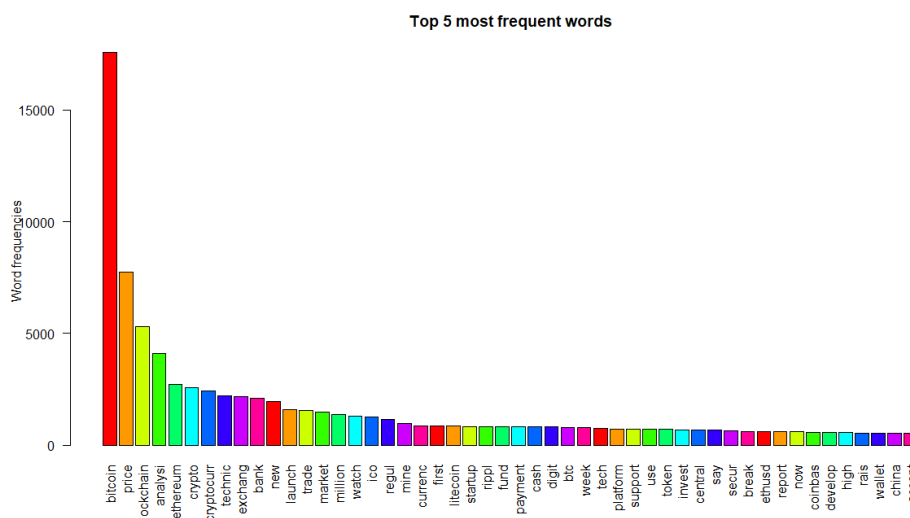


Figure 6: Most occurred words in data.

Based on the most frequent words, a word cloud is generated to visualize the keywords found within the text in terms of frequencies. The comparison cloud in Figure 7 presents the frequent topics headlined throughout the year. Each color is representing a year. The word cloud appears that Bitcoin is the main keyword in the headlines of 2013. According to Forbes, 2013 is said to be the year of Bitcoin as the cryptocurrency prices trade higher than gold in the financial market⁶. The year is also the beginning of cryptocurrency where most investors started to pay attention to digital currency. Ethereum appeared to be one of the main terms in 2016, indicating the digital currency is the popular currency other than Bitcoin. Furthermore, the term ‘blockchain’ also appeared on many news headlines in 2016.

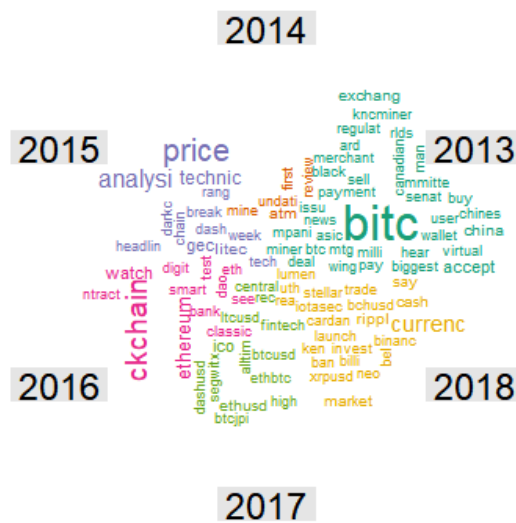


Figure 7: Comparative word cloud.

Figure 8 shows the distribution of cryptocurrency headlines throughout the five years. It is noteworthy that the trend of cryptocurrency is going upward, demonstrating highly concerned with news media surrounding the topic of cryptocurrency. The number of headlines increased from 2,000 in 2013 to approximately 10,000 in 2018.

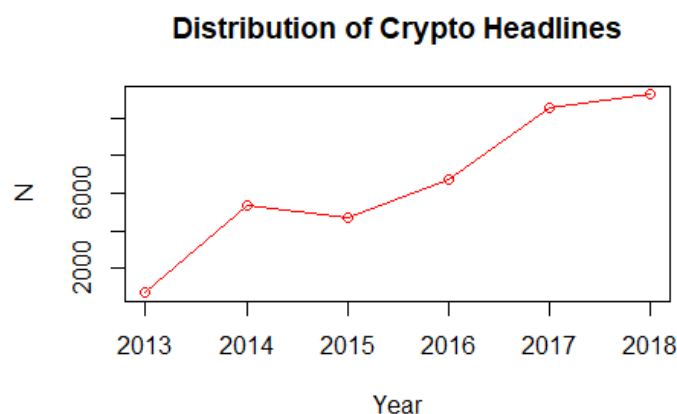


Figure 8: Distribution of Crypto headlines.

⁶ <https://www.forbes.com/sites/kitconews/2013/12/10/2013-year-of-the-bitcoin>

6.2 Sentiment Lexicon

Table 1 2: Lists of sentiment for general model

Sentiment	N
Negative	10,621
Positive	8,870

The sentiment result with Bing approach shows the negative sentiment in the news headlines is more than positive sentiment. This could be because of the lists of negative words (4,873) in Bing dictionary is more than positive (2,006) words. Polarity scores are measured using equation (4) and diagrams are plotted using ggplot to demonstrate the changes of sentiment over time. Positive sentiment is also extracted using equation (5) as we are interested to know the positive changes in headlines. Despite the overall sentiment trend from 2013 to 2018 is negative, but the sentiment is decreasing year by year. Based on the result, we can tell that the individuals started to react positively towards cryptocurrency. In addition to that, the plot on the right in Figure 9 showed an upward trend in terms of positive polarity score. According to Figure 10, we once know that ‘support’ appeared to be the most occurrences words in positive sentiment while ‘break’ shows the most negative words.

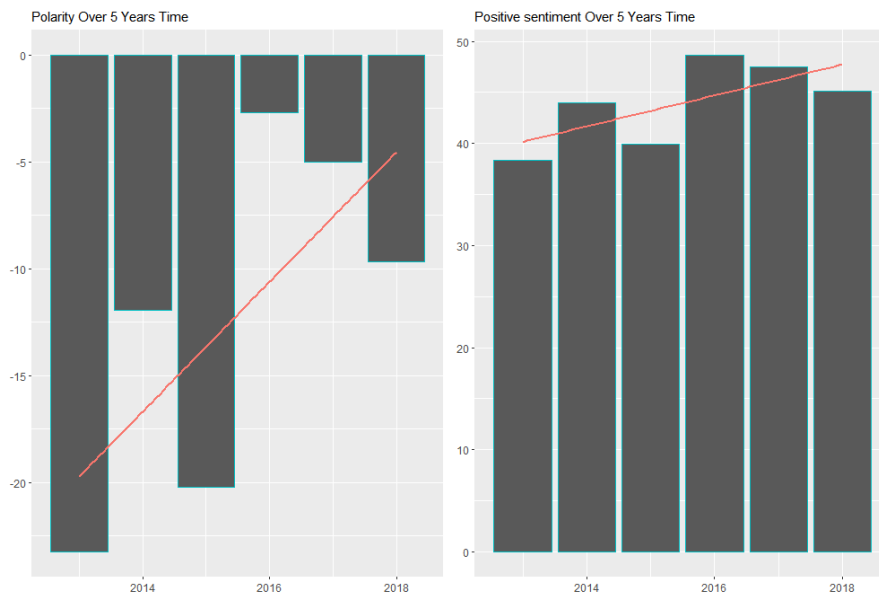


Figure 9: Overall polarity and positive polarity.

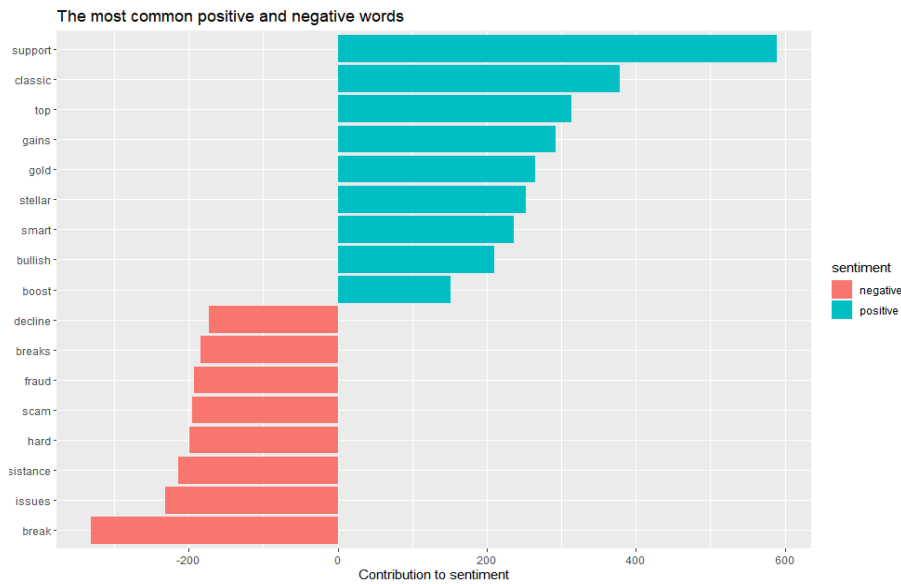


Figure 10: The top sentiment words general model.

TF-IDF feature is used as an input to SVM for training and the result (Table 3) shows the value of both sentiments has reduced compared to the previous general model. Figure 11 presents the top sentiment words with common words. We may notice that there is a minor change in the word's sentiment. The word 'hard' in the negative contribution moved from the fourth position in model 1 to the second position in model 2 after retaining only important words.

Table1 3: Lists of sentiment for TF-IDF model

Sentiment	N
Negative	1,197
Positive	647

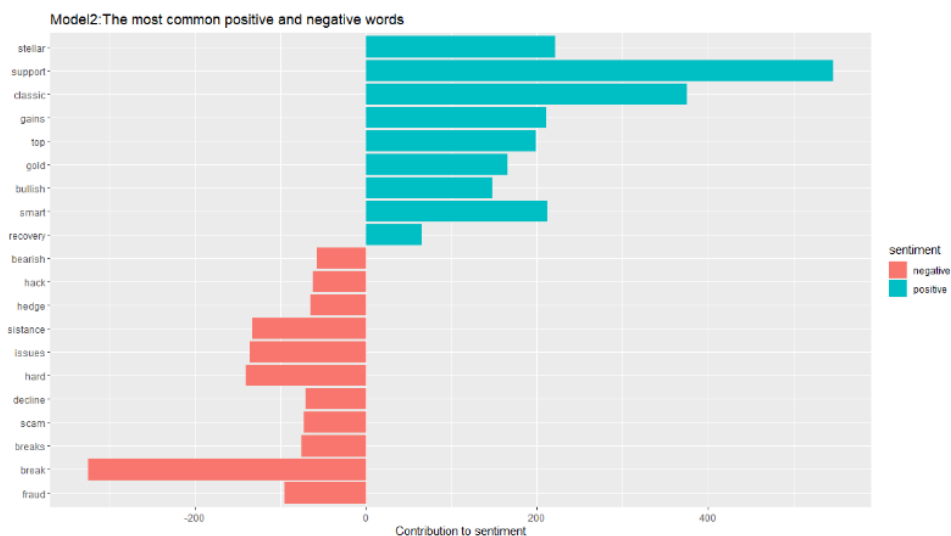


Figure 11: The top sentiment for TF-IDF model.

6.3 Support Vector Machine

Table 4: Accuracy of two models

	Accuracy
General model	0.5504
TF-IDF model	0.6143

SVM is applied after the polarity is measured in the previous section to evaluate the sentiment result in the Bing lexicon. In the paper, two models are used for modelling: general model with entire sentiment words, TF-IDF model with the assigned weight on each word. Each model is trained with a linear kernel function using sentiment as the response variable consisting of two classes ‘negative’ and ‘positive’.

By using the summary command, there are no training errors and the cost equal to 10 is used in the kernel function. For general model, there is 14,238 number of supporting vectors. The accuracy of test data in model 1 is 55% with a zero kappa. From the test model prediction, 2,145 of negative sentiment is correctly predicted. The second model performed well with an accuracy of 61% and there are 2,745 supporting vectors, 1,374 in one class, and 1,371 in another class. The caret function is used to present the performance of the train and test model. There is an average improvement of 6% when compared to the result obtained by model 1. The zero value of Kappa indicates the predictions mostly fall on negative class. Based on the test model prediction, 559 negative sentiments are truly predicted, and 351 positive sentiments are falsely predicted.

6.4 Discussion

In this segment, we measured the score of all headlines consist of positive and negative which represented in the ggplot (Figure 9). The trend of five years of data indicated a negative sentiment in the crypto headlines by various online media publications. The possible reason for this is due to the imbalance sentiment classification in Bing’s dictionary that leads to the output obtained. On the other way, we extract and measure only the positive polarity and the plot indicated an upward trend in which we can say that the attitude of media publication towards cryptocurrency is improving over time.

In addition, the year 2016 acquired the lowest negative polarity score and this could be because of an annual gain of 54% in cryptocurrency during the year⁷. As highlighted above, the TF-IDF model outperformed the general lexicon model in the SVM, demonstrating a better accuracy of classification. The finding is consistent with Nasim *et al.*, (2017) who also achieve the best performance in using TF-IDF and sentiment lexicon on student feedback. The paper concluded to this by comparing to four approaches including the general lexicon-based approach and obtained 2% of improvement in the result. Haddi *et al.*, (2013) also achieved

⁷ <https://www.businessinsider.com/heres-why-bitcoin-boomed-in-2016-2016-12?r=US&IR=T>

desirable results in applying TF-IDF matrix for online movie reviews. The paper used three types of weighting features: TF-IDF, FF, and FP in the pre-processing to reduce the noise in the text and ran the SVM classifier with each feature. The IF-IDF delivered the accuracy of 93.5% higher than the other two models.

7 Conclusion and Future Work

In the sentiment analysis on the area of cryptocurrency from online new platforms did successfully answered the research question mentioned in the introduction part. By applying the Bing approach, we were able to identify that though it was a negative sentiment in the headlines of cryptocurrencies over the past five years, yet the negative sentiment is reducing year by year, proving that sentiment towards the crypto market is on a rise. This indicated that cryptocurrency is gaining more confidence from investors and news media platforms. By excluding Bitcoin in the text, the paper also found that Ethereum, Litecoin, and Ripple are gaining a portion in the market. These digital currencies are included under the list of top 40 most frequently occurred words. This could be a valuable insight for new investors to have an idea of the trend of cryptocurrency.

From the SVM algorithm, we can know that most of the words ranked under negative class. This could be assumed due to the huge gap between the sentiment variable. There are imbalance classes in the data where more negative words are contained in the Bing list of words than positive. By comparing the general model and TD-IDF model, the second model has improved the performance in terms of accuracy and hence outperformed the general model. Overall, the attitude of online news publication is getting better in the evidence from the number of published crypto-related news data increased from 2013 to 2018. The positive polarity score is moving upward in the line graph as people are getting more familiar with cryptocurrency.

In future work, improvement can be done by considering emoticon to analyze the text document. Currently, we evaluate the headlines based on binary classification. It would be interesting to see how other emoticon features reflect the sentiment result. We are also interested to experiment with more sophisticated features in the future.

References

- Asghar, M. Z. et al., 2017. Lexicon-enhanced sentiment analysis framework using rule-based classification scheme. *PLoS ONE*, 12(2), pp. 1-22.
- Augustyniak, L., Szymanski, P., Kajdanowicz, T. & Tuligłowicz, W., 2016. Comprehensive Study on Lexicon-based Ensemble Classification Sentiment Analysis. *Entropy*, 18(4), pp. 1-29.
- Chu, J., Chan, S., Nadarajah, S. & Osterrieder, J., 2017. GARCH Modelling of Cryptocurrencies. *Journal of Risk and Financial Management*, 10(4), pp. 1-17.

- Ciaian, P., Rajcaniova, M. & Kanacs, d., 2016. The digital agenda of virtual currencies: Can BitCoin become a global currency?. *Information Systems and e-Business Management*, 14(4), pp. 883-919.
- David, L. K. C., Guo, L. & Wang, Y., 2017. Cryptocurrency: A new investment opportunity. *The Journal of Alternative Investments*, 20(3), pp. 16-40.
- Dhaoui, C., Webster, C. M. & Tan, L. P., 2017. Social media sentiment analysis: lexicon versus machine learning. *Journal of Consumer Marketing.*, 34(6), pp. 480- 488.
- Donmez, C. C., Dereli, A. F., Horasan, M. B. & Yildiz, C., 2020. An analysis of Evolutionary Cryptocurrency Market Dynamics. *Electronic Journal of Social Sciences*, 19(74), pp. 611-629.
- Dulău, T.-M. & Dulău, M., 2019. Cryptocurrency–Sentiment Analysis in Social Media. *Acta Marisiensis. Seria Technologica*, 16(2), pp. 1-6.
- Ersahin, B., Özlem, A., Kilinc, D. & Deniz, E., 2019. A hybrid sentiment analysis method for Turkish. *Turkish Journal of Electrical Engineering & Computer Sciences*, 27(3), pp. 1780-1793.
- Gan, B. A., Bird, R. & Yeung, D., 2020. Sensitivity to sentiment: News vs social media. *International Review of Financial Analysis*, Volume 67, pp. 1-17.
- Gidea, M. et al., 2020. Topological recognition of critical transitions in time series of cryptocurrencies. *Physica A: Statistical Mechanics and its Applications*, Volume 548, p. 123843.
- Glenski, M., Saldanha, E. & Volkova, S., 2019. *Characterizing speed and scale of cryptocurrency discussion spread on reddit*. New York, United States, In The World Wide Web Conference.
- Gurriba, I., Kwehb, Q. L., Nouranic, M. & Ting, I. W. K., 2019. Are Cryptocurrencies Affected by Their Asset Class Movements or News Announcements?. *Malaysian Journal of Economic Studies* , 56(2), pp. 201-225.
- Haddi, E., Liu, X. & Shi, Y., 2013. The Role of Text Pre-processing in Sentiment Analysis. *Information Technology and Quantitative Management*, Volume 17, pp. 26-32.
- Han, H. et al., 2018. Improving the performance of lexicon-based review sentiment analysis method by reducing additional introduced sentiment bias. *PLOS ONE*, 13(8), pp. 1-11.
- Im, T. L., San, P. W., On, C. K. & Anthony, P., 2015. *Rule-based Sentiment Analysis for Financial News*. Kowloon, China, IEEE International Conference on Systems, Man, and Cybernetics.
- Kang, K., Choo, J. & Kim, Y., 2019. Whose Opinion Matters? Analyzing Relationships Between Bitcoin Prices and User Groups in Online Community. *Social Science Computer Review*, Volume , pp. 1-17.
- Kaur, B. & Kumari, N., 2016. A Hybrid Approach to Sentiment Analysis of Technical Article Reviews. *International Journal of Education and Management Engineering*, 6(6), pp. 1-11.

- Kim, Y. B. et al., 2016. Predicting Fluctuations in Cryptocurrency Transactions Based on User Comments and Replies. *PloS one*, 11(8), pp. 1-17.
- Kim, Y. B. et al., 2017. When Bitcoin encounters information in an online forum: Using text mining to analyse user opinions and predict value fluctuation. *PloS one*, 12(5), pp. 1-14.
- Lamon, C., Nielsen, E. & Redondo, E., 2017. Cryptocurrency Price Prediction Using News and Social Media Sentiment. *SMU Data Sci. Rev*, 1(3), pp. 1-22.
- Liu, B., Hu, M. & Cheng, J., 2005. *Opinion observer: analyzing and comparing opinions on the Web*. s.l., In Proceedings of the 14th international Conference on World Wide Web.
- Mohammad, S. M. & Turney, P. D., 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3), pp. 436-465.
- Nakamoto, S., 2009. Bitcoin: A Peer-to-Peer Electronic Cash System.
- Narman, H. S. & Uulu, A. D., 2020. *Impacts of Positive and Negative Comments of Social Media Users to Cryptocurrency*. Big Island, HI, USA, IEEE, pp. 187-192.
- Nasim, Z., Rajput, Q. & Haider, S., 2017. *Sentiment analysis of student feedback using machine learning and lexicon based approaches*. Langkawi, Malaysia, International Conference on Research and Innovation in Information Systems (ICRIIS).
- Nielsen, F. Å., 2011. *A new ANEW: Evaluation of a word list for sentiment analysis in microblogs*. s.l., Proceedings of the ESWC.
- Park, H. W. & Lee, Y., 2019. How Are Twitter Activities Related to Top Cryptocurrencies' Performance? Evidence From Social Media Network and Sentiment Analysis. *Drustvena Istrazivanja; Zagreb*, 28(3), pp. 435-460.
- Piatetsky-Shapiro, G., 1991. Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop. *AI Magazine*, 11(5), pp. 68-70.
- Rognone, L., Hyde, S. & Zhang, S., 2020. News sentiment in the cryptocurrency market: An empirical comparison with Forex. *International Review of Financial Analysis*, Volume 69, pp. 1-17.
- Sajter, D., 2019. Time-Series Analysis of the Most Common Cryptocurrencies. *Ekonomski misao i praksa*, 13(1), pp. 267-282.
- Steinert, L. & Herff, C., 2018. Predicting altcoin returns using social media. *PLoS ONE*, 13(12), pp. 1-12.
- Taboada, M. et al., 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2), pp. 267-307.
- Valencia, F., Gómez-Espinosa, A. & Valdés-Aguirre, B., 2019. Price Movement Prediction of Cryptocurrencies Using Sentiment Analysis and Machine Learning. *Entropy*, 21(6), p. 589.
- Wooley, S., Edmonds, A., Bagavathi, A. & Krishnan, S., 2019. *Extracting Cryptocurrency Price Movements From the Reddit Network Sentiment*. Boca Raton, FL, USA, IEEE International Conference On Machine Learning And Applications (ICMLA).

Yu, T. & Nwet, K. T., 2020. *Comparing SVM and KNN Algorithms for Myanmar News Sentiment Analysis System*. New York, United States, Association for Computing Machinery.

Zainuddin, N., Selamat, A. & Ibrahim, R., 2017. Hybrid sentiment classification on twitter aspect-based sentiment analysis. *Applied Intelligence* , 48(5), pp. 1218-1232.