Language Resources and Evaluation (2025) 59:807–842
https://doi.org/10.1007/s10579-024-09733-z

ORIGINAL PAPER

Dataset on sentiment-based cryptocurrency-related news and tweets in English and Malay language

Nur Azmina Mohamad Zamani[1,2] · Norhaslinda Kamaruddin[3] · Ahmad Muhyiddin B. Yusof[4]

Accepted: 6 March 2024 / Published online: 20 May 2024

Abstract

Cryptocurrency trading is becoming popular due to its profitable investment and has led to worldwide involvement in buying and selling cryptocurrency assets. Sentiments expressed by cryptocurrency enthusiasts toward some news via social media or other online platforms may affect the cryptocurrency market activities. Thus, it has become a challenge to determine the level of positivity or negativity (regression) inhibiting the texts than simply classifying the sentiment into categorical classes. Regression offers more detailed information than a simple classification which can be robust to noisy data as they consider the entire range of possible target values. On the contrary, classification can lead to biased models due to imbalanced dataset and tend to cause overfitting. Hence, this work emphasises in creating sentiment-based cryptocurrency-related corpora in English and Malay focusing on Bitcoin and Ethereum. The data was collected from January to December 2021 from the publicly available news online and tweets from Twitter in English and Malay. The dataset contains a total of 29,694 instances comprised of 5694 news data and 24,000 tweets data. During the annotation process, the annotators are trained until Krippendorf's alpha agreement of above 60% is achieved since it is considered an applicable benckmark due to the annotation complexity. The corpora is available on Github for cryptocurrency-related experiments using various machine learning or deep learning models to study English and Malay sentiments effect on the global market, particularly the Malaysian market and can be extended for further analysis for Bitcoin and Ethereum market volatile nature.

Keywords Sentiment corpora · Cryptocurrency · Dataset · Malay texts

Extended author information available on the last page of the article

1 Introduction

Cryptocurrency is a new form of digital currency that exploits blockchain technology and cryptographic functions, “mined” by any individuals using own resources such as computers or any electronic devices without being ruled by any financial institutions or government authorities (Daskalakis & Georgitseas, 2020; Goleman, 2018; Pintelas et al., 2020). Bitcoin (BTC) is the initial well-known cryptocurrency coin that was introduced by Nakamoto (2008) and made public in 2009. According to Forbes news in October 2023, Bitcoin has the highest market capitalisation with the value of about £443 billion (Hooson & Pratt, 2023). Other cryptocurrency coin (also known as altcoin) are also becoming widespread among traders such as Ethereum (ETH) with the current market capitalisation value of £163 billion which comes second after Bitcoin (Hooson & Pratt, 2023). Due to a higher chance of gaining profits through cryptocurrency trading and investments, it has garnered significant interest among researchers, traders and cryptocurrency enthusiasts over the years to investigate the causes and effects of the cryptocurrency market activities (Kang et al., 2022). Thus, explorations on sentiment affecting crucially on the cryptocurrency market behaviour began to be conducted actively via a variety of experiments (Balfagih & Keselj, 2019; Cerda, 2021; Chin & Omar, 2020; Edgari et al., 2022).

A cryptocurrency influencer like Elon Musk is capable of dictating the directions of the cryptocurrency prices, mainly during some news postings on social media platform such as Twitter that contains positive or negative sentiment expressions about some cryptocurrency coin(s). Such events may have persuasive impact on cryptocurrency traders and investors to sell, buy or hold their cryptocurrency asset(s) (Pintelas et al., 2020). Hence, computational models for sentiment analysis are required to employ sentiment features from massive number of texts in social media postings and news headlines regarding cryptocurrency issues.

The current available sentiment-based dataset regarding cryptocurrency is limited and mainly in English (Garg et al., 2021; Lamon et al., 2017; Mohapatra et al., 2019; Seroyizhko et al., 2022) and to the best of our knowledge, there is very limited Malay corpus on sentiment-based cryptocurrency available. Nevertheless, our work on sentiment analysis using our own curated English and Malay news corpora have been published (Zamani et al., 2022a, 2022b; Zamani & Kamaruddin, 2023). Therefore, this paper aims to annotate a total of 29,797 text documents to construct a total of eight sentiment corpora:

1) BTC English News
2) ETH English News
3) BTC English Tweets
4) ETH English Tweets
5) BTC Malay News
6) ETH Malay News
7) BTC Malay Tweets

8) ETH Malay Tweets

The creation of these corpora involves training process among annotators, whereby a rubrics and codebook will be given as a guide to label the sentiment texts accordingly. This process is challenging since the labelling is considered to be quite subjective as the range of sentiment scores is between − 1 (very negative) to + 1 (very positive), instead of a general classification scores such as + 1 (positive), 0 (neutral), or − 1 (negative) (Cerda, 2021; Chen et al., 2019; Chin & Omar, 2020). The 3-class (positive, negative, and neutral) sentiment annotation is widely adopted because typically, a sentence comprises of a single sentiment (Hartmann et al., 2023; Riccosan & Saputra, 2023). However, this is not usually the case since a statement may comprise multiple polarity (mixed sentiments) (Hu & Liu, 2004). To simplify, the categorical polarity may not reflect the actual overall sentiment intensity (Chawla et al., 2004; Taboada et al., 2011). For instance, the moderately positive sentiment is not similar to the least positive sentiment and yet both statements are represented as + 1 in the 3-class sentiment annotation. The regression labelling overcome such issue by signifying a different weight value in positive category. Such approach allows more detailed analysis which is important to capture relationships between instances for sentiment analysis (Nandwani & Verma, 2021). The nuances of the sentiments are captured, and these subtle differences can offer better sentiment understanding (Yang et al., 2021).

Based on existing research studies, an automated labelling using Valence Aware Dictionary for Sentiment Reasoning (VADER) or TextBlob is used to obtain the sentiment scores (Serafini et al., 2020; Steinert & Herff, 2018; Valencia et al., 2019). However, these tools only support English texts, thus, motivates our study to create our own sentiment corpora specifically on Bitcoin and Ethereum cryptocurrency in English and Malay. Our corpora creation involves processes such as data extraction: English online news using NewsAPI, Malay online news using Parsehub, and Twitter (English and Malay) using SNScrape, text pre-processing, and annotation. The corpora will then be published on GitHub to benefit other research studies in widening the research area on cryptocurrency, especially in non-English texts. In addition, this paper presented several machine learning models regression to illustrate the performance of the curated datasets.

This paper is organised as follows: Sect. 2 presents the related works on cryptocurrency datasets, followed by the description of dataset in Sect. 3. Section 4 explains the experimental setup, then the result and discussion is presented in Sect. 5. Lastly, the concluding remark and future work are stated in Sect. 6.

2 Related works

This section presents how the labelling was performed in prior research studies on cryptocurrency price prediction using sentiment analysis. Several commonly applied methods are using automated tools such as VADER, Textblob, and other Python library packages. Dictionaries are also being utilised to obtain the polarity scores

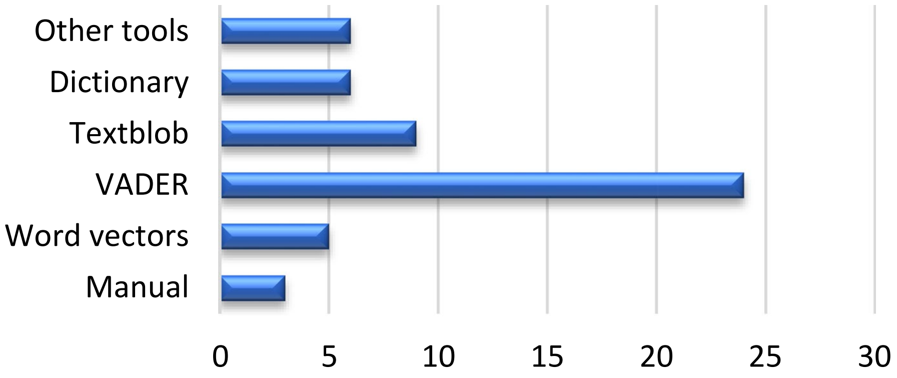


Fig. 1 Statistics on the labelling methods used in existing research studies

of the text. In addition, manual annotation is also being performed by some current research studies. Figure 1 shows the statistics of labelling methods applied in existing research studies from the year 2017 to 2022 that performed sentiment analysis for cryptocurrency price prediction.

Based on Fig. 1, it is observed that Valence Aware Dictionary for Sentiment Reasoning (VADER) was commonly used for data labelling in the current research studies on cryptocurrency price prediction with sentiment analysis, followed by Textblob and other methods such as the application of dictionary of words, manual annotation, and taking the word vectors produced by some pretrained word embedding tool. VADER is more popular to be used to automatically label sentiments compared to other automated labelling tools because its lexicon is a gold-standard quality that has been validated by humans, and provides a fast labelling sentiment data with good performance in labelling social media data (Bonta et al., 2019). Although automatic approach in annotating sentiment data needs less time to perform, it is more prone to produce some errors compared to manual annotation approach (Manaf & Nordin, 2009). Despite manual annotation time-consuming, complicated, and subjective behaviours, this approach offers a better-quality labelling on the data by humans (Manaf & Nordin, 2009). This is because, human annotators can understand and interpret context with their domain expertise that can annotate more precisely than automatic methods by handling ambiguity through a human's judgement (D'Orazio et al., 2022). Hence, manual annotation was selected for this study to produce a good quality sentiment-labelled corpora.

Jahjah and Rajab (2020) used VADER to label 25,746 tweets on Bitcoin, then a word embedding called Global Vectors for Word Representation (GloVe) was implemented and fed into a hybrid deep learning sentiment model of Convolutional Neural Network (CNN)—Long Short-Term Memory (LSTM), that yielded an accuracy of 88.7% for the sentiment analysis evaluation. Whilst, other related studies from the 45% of the statistics in Fig. 1 utilising VADER as the labelling technique did not perform sentiment model evaluation, as they directly feed the computed sentiment scores into the cryptocurrency price prediction model (Aggarwal et al., 2019; Critien et al., 2022; Edgari et al., 2022; Ibrahim, 2021;

Loginova et al., 2021; Maqsood et al., 2022; Mohapatra et al., 2019; Oikonomopoulos et al., 2022; Parekh et al., 2022; Pillai et al., 2021; Prajapati, 2020; Salač, 2019; Sattarov et al., 2020; Schulte & Eggert, 2021; Serafini et al., 2020; Steinert & Herff, 2018; Stenqvist & Lönnö, 2017; Wołk, 2019; Wooley et al., 2019; Yao et al., 2019).

Textblob is the next most frequently employed tool for data labelling for cryptocurrency-related sentiment analysis (Agarwal et al., 2021; Alghamdi et al., 2022; Hitam et al., 2021; Jain et al., 2018; Kilimci, 2020; Loginova et al., 2021; Pathak & Kakkar, 2020; Prajapati, 2020; Raju & Tarif, 2020). It is found that by using Textblob with FastText word embedding by Kilimci (2020) on 17,629 tweets has produced a result of 89.13% accuracy. In addition, Alghamdi et al. (2022) has tested the utilization of Textblob using the Support Vector Machine (SVM) classifier on 75,995 tweets (Bitcoin and Ethereum) and obtained an accuracy of 93.95% and 95.59% for Bitcoin and Ethereum respectively.

Various type of dictionaries was used in prior studies to label the sentiment data. Mai et al. (2018) and Gurdgiev and O'Loughlin (2020) have applied Loughran & McDonald dictionary (Loughran & Mcdonald, 2011, 2014) on tweets and forum accordingly. In addition, Mai et al. (2018) also used Thomson Reuters News Analytics (TRNA) database to produce news sentiment scores. From their investigation, it is confirmed that sentiment has an important link to the cryptocurrency market activities. On the other hand, Hasan et al. (2022) employed Alex Davies wordlist to classify 10,260 tweets into positive, negative, and neutral but no sentiment analysis evaluation is reported in this work. A cryptocurrency-specific lexicon dictionary consisting of positive and negative terms was applied on a total of 1,533,975 Stocktwits and 1,392,587 Reddit posts, where the sentiment scores are computed using the designated formula stated in Chen et al. (2019) producing an accuracy of 86%. Meanwhile, NTUSD-Fin dictionary on LSTM was employed by Chin and Omar (2020) on 15,295 online news with an accuracy of 65.83% for the sentiment analysis result. Hence, this demonstrates that a larger sized corpus with some machine learning model plays an important part in the performance of a sentiment model.

Moreover, Cerda (2021) performed two types of analysis: (1) sentiment analysis, (2) stance detection in which a total of 9146 tweets were classified. For sentiment analysis, SentiWordNet was utilised and achieved F1-score of 47%, whereas for stance detection, the data was labelled with either support, against, or none. Bidirectional Encoder Representations from Transformers (BERT) classifier was implemented to analyse the performance of stance detection and obtained F1-score of 67%. Similarly, BERT was used in Passalis et al. (2022) work to classify the data (i.e., Reddit, online news, tweets, and Telegram posts), with the sentiment analysis accuracy of 68%. Besides using the state-of-the-art (SOTA) language model, traditional word embedding such as Word2Vec was also applied for sentiment classification (Mohanty et al., 2018; Nasekin & Chen, 2020). Nasekin and Chen (2020) applied Word2Vec on 1,220,728 Stocktwits and fed the text representation into LSTM model achieving an F1-score of 85%. Interestingly, a sentiment regression task was performed by Attila (2017) on tweets using GloVe word embedding and CNN model, thus, the sentiment prediction resulted in a mean squared error (MSE) of 0.927.

Other automated data labelling was also performed using tools such as Pattern Python package (Galeshchuk et al., 2018), Keras library Python package (Inamdar et al., 2019), Flair (Prajapati, 2020), Weighted Event Sentiment Score (WESS) (Rognone et al., 2020), Python labelling of positive, negative, and neutral (Shahzad et al., 2021), and Joint/Sentiment Topic (JST) model (Loginova et al., 2021). Some related work also conducted manual data annotation, especially on non-English texts. Lamon et al. (2017) annotated an approximate of 3600 news headlines and 10,000 tweets manually using daily coin price to perform binary classification of 0 (decrease in price) and 1 (increase in price) for the data. Similar technique of relying on the cryptocurrency price rise and drop was done by Vo et al. (2019) where the data was labelled with 1 (positive), 0 (neutral), or − 1 (negative). Subsequently, manual classification (i.e., positive, negative, irrelevant/neutral) conducted by Pant et al. (2018) on 7454 tweets obtained an accuracy of 78.49% with the implementation of Bag-of-Word technique with Random Forest Classifier for sentiment analysis. We have also published a work on cryptocurrency-related sentiment regression (Zamani, Liew, et al., 2022) that performs manual annotation on both English and Malay news headlines (Bitcoin and Ethereum) for corpora creation. A total of 5697 news headlines were manually annotated by three annotators, where each news headline was labelled with scores between − 1 (very negative) and 1 (very positive). The supervised datasets were then being experimented using Generalized Autoregressive Pretraining for Language (XLNet) – Gated Recurrent Unit (GRU) for sentiment regression performance evaluation. The results reported are as follows: (1) Adjusted $R^2=0.654$ (Bitcoin English news), (2) Adjusted $R^2=0.607$ (Ethereum English news), (3) Adjusted $R^2=0.\ 428$ (Bitcoin Malay news), and (4) Adjusted $R^2=0.599$ (Ethereum Malay news). Table 1 shows the annotation methods applied in prior related research studies with its sentiment model's performance.

Based on Table 1, it can be observed that most large-sized dataset tend to be labelled using automated tools, while manual approach was conducted for smaller dataset (less than 10,000 instances). However, it can be perceived that even though manual annotation on smaller dataset gave slightly lower accuracy than some of the larger dataset with automated labelling, it still proves to be an efficient approach referring to a comparison between Pant et al. (2018) with manual annotation on 7454 tweets achieving 10.49% higher in the accuracy than in Passalis et al. (2022) using BERT to automatically labelled 200,000 instances. Moreover, Jahjah and Rajab (2020) used the popularly used VADER to label 25,746 data and achieved an accuracy of 88.7% whereas other existing studies applying manual annotations shown the accuracy results of between 67% to 78.49%. This infers that manual annotations can still be a superior approach since small dataset with good quality data labelling can provide good comparable accuracy.

3 Dataset creation

Malaysia has begun its involvement in cryptocurrency since 2019 with the approval of the Malaysian government following its laws and regulations (Farhana & Muthaiyah, 2022; Sukumaran et al., 2022), thus, "Luno" is being used legally as the

Table 1 Summary of sentiment analysis performance using various annotation approach

Author	Size of Dataset	Annotation Method	Model	Results
Jahjah and Rajab (2020)	25,746	VADER	GloVe + CNN-LSTM	Accuracy = 88.7%
Kilimci (2020)	17,629	TextBlob	FastText	Accuracy = 89.13%
Alghamdi et al. (2022)	75,995	TextBlob	SVM	Accuracy = 93.95% (BTC), 95.59% (ETH)
Chen et al. (2019)	More than a milion	Crypto-specific dictionary	Own model	Accuracy = 86%
Chin and Omar (2020)	15,295	NTUSD-Fin dictionary	LSTM	Accuracy = 65.83%
Cerda (2021)	9146	Manual	BERT + XGBoost	F1-score = 67%
Passalis et al. (2022)	200,000	BERT	CNN	Accuracy = 68%
Nasekin and Chen (2020)	1,220,728	Word2vec	LSTM	F1-score = 85%
Attila (2017)	More than a million	GloVe	CNN	MSE = 0.927
Pant et al. (2018)	7454	Manual	BOW + Random Forest	Accuracy = 78.49%
Zamani, Liew, et al. (2022)	5697	Manual	XLNet-GRU	1) Adjusted $R^2 = 0.654$ (BTC English news), 2) Adjusted $R^2 = 0.607$ (ETH English news), 3) Adjusted $R^2 = 0.428$ (BTC Malay news), 4) Adjusted $R^2 = 0.599$ (ETH Malay news)

cryptocurrency exchange platform for the users in Malaysia to perform the cryptocurrency transactions (Yusof et al., 2021). Since Malaysian users tend to express some strong sentiments in their mother tongue language which is Malay in addition to English words, there is a significant need to explore both English and Malay sentiment on the Malaysia market, which may also affect the financial market globally. The duration of dataset collection is a year (1 January 2021–31 December 2021) during the chaotic economic time of COVID-19 season, where most activities were actively done online due to the Movement Control Order (MCO) restricting people from doing outdoor activities (Ahmad et al., 2023; Barbaglia et al., 2022; Luo, 2020). The data collection throughout the year 2021 is significant because there was some economic chaos amids the pandemic season of COVID-19 that leads to unstable or extreme market volatility and price fluctuation at some point (Lahmiri & Bekiros, 2020). Moreover, during the World Health Organization (WHO) announcement of the global pandemic, it is found that the cryptocurrency market liquidity increased, both sharply and significantly (Corbet et al., 2022). This indicates that the 2021 sentiment data on cryptocurrency can provide a more significant pattern to be investigated in the cryptocurrency market behaviour.

First, we extract the news headlines using NewsAPI (English news) and Parsehub (Malay news), and tweets using Twitter SNScrape (English and Malay tweets). NewsAPI is a Python library applied in the implementation to get English news online from over 30,000 sources (e.g., Cointelegraph, CNN, Business Insider, Forbes, etc.) all over the world created by Lisivick (2017), while Parsehub is a user friendly tool that can be downloaded from https://www.parsehub.com, where it extracts the contents of webpages. Since there are no Malay news extraction Python library available, Parsehub was being utilised to gather all the Malay online news from various sources (i.e., Intraday.my, Berita Harian, Astro Awani, etc.). Intraday.my is a cryptocurrency-related online news that published in Malay language and this is one of the main source to gather the data for this study as it contains voluminous information on cryptocurrency compared to the other general local news articles in Malaysia. SNScrape is a scraper for social networking services (SNS) using user information, hashtags, or searches, then produce the relevant outputs such as tweets from Twitter (SNScrape, 2018). It was also being implemented using Python programming to obtain the extracted results. Hence, there are a total of more than 2000 news headlines and over 24 million tweets being extracted. Then, we split the data into a total of eight separate corpus.

From the raw data gathered for the news headlines, the data were splitted manually into four different corpus: (1) English Bitcoin news, (2) English Ethereum news, (3) Malay Bitcoin news, and (4) Malay Ethereum news. Whilst, for Twitter scraping, we used the following queries to extract tweets on Bitcoin and Ethereum: “bitcoin”, “btc”, “ethereum”, “eth” since these cryptocurrency coins are the current top cryptocurrency with high market capitalization, indicating the large volume of their traders. These two cryptocurrency coins were also being used as the case study in most of the existing research studies (Alghamdi et al., 2022; Chowdhury et al., 2020; Hasan et al., 2022; Loginova et al., 2021; Ortu et al., 2022). The data were then stored into four different corpus: (1) English Bitcoin tweets, (2) English Ethereum tweets, (3) Malay Bitcoin tweets, and (4)

Malay Ethereum tweets. The news headlines and tweets were filtered manually by removing all the non-English texts for English news corpora and removing non-Malay texts for Malay corpora before going through pre-processing. All extracted news headlines are taken for manual annotation. However, due to the very large amount of raw tweets extracted and limited time, only 6000 tweets for each tweets corpus were selected for the manual annotation process. Based on the related reviewed literature on the number of dataset for manual annotation, majority of the studies tend to annotate less than 10,000 number of data manually. Therefore, from a total of over two millions raw data for each English and Malay tweets, 6000 tweets is sufficient for manual annotation after removing noisy raw data and non-English or non-Malay texts in the respective corpus. In addition, the 6000 instances were the sum obtained from 500 instances in each month (total 12 months) to get a balanced dataset for tweets. This also aligns with the manual annotation performed by Liu et al. (2023) on 6500 tweets achieving an accuracy of about 72%. Table 2 shows the number of data selected for each corpus.

The chosen number of data to be annotated as shown in Table 1 were based on the prior research studies (Cerda, 2021; Pant et al., 2018; Rognone et al., 2020) performing manual annotations on sentiment data for cryptocurrency price prediction topic. Figure 2 illustrates the summary of the existing related works that perform manual sentiment labelling.

Based on Fig. 2, other prior related studies found to perform manual annotation had annotated between 3108 and 9146 number of instances in English languages. There were not many cryptocurrency-related studies performed manual sentiment annotations especially in languages other than English. Hence, this study has conducted a slighly larger amount of data to be annotated since it has been shown in some existing studies that a larger number of data may lead to a better performance and small datasets may trigger over-fitting of the sentiment model (Althnian et al., 2021). The corpora creation framework is shown in Fig. 3.

By using the collected data with the filtration of non-English and non-Malay texts, the architecture proceeds with the text pre-processing where unknown symbols, insignificant special characters, and hyperlinks are removed to produce cleaner texts for annotation process. This also includes the emoticons handling for tweets data.

Table 2 Number of data selected for news headlines and tweets corpora

Corpus	Cryptocurrency	English	Malay
News headlines	Bitcoin	1518	1520
	Ethereum	1204	1452
Tweets	Bitcoin	6000	6000
	Ethereum	6000	6000

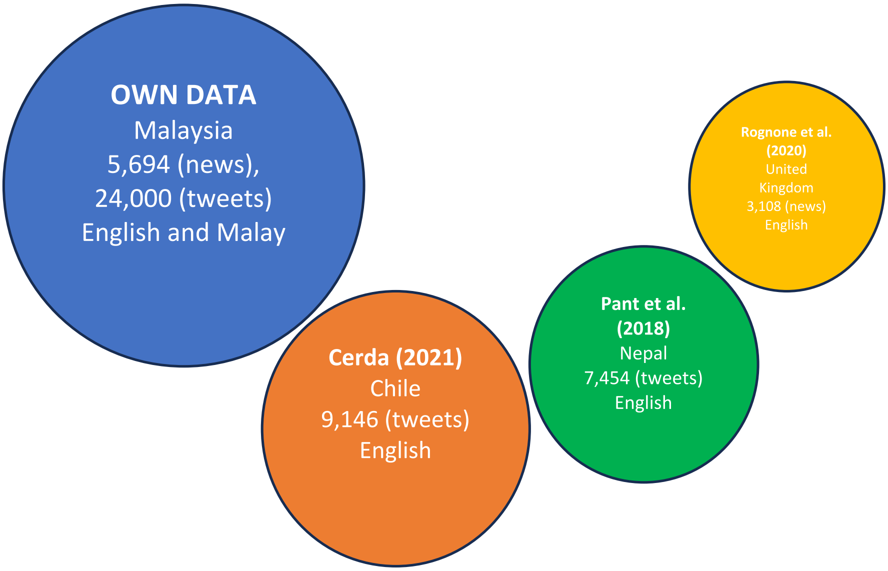


Fig. 2 Number of data for manual labelling in existing research studies

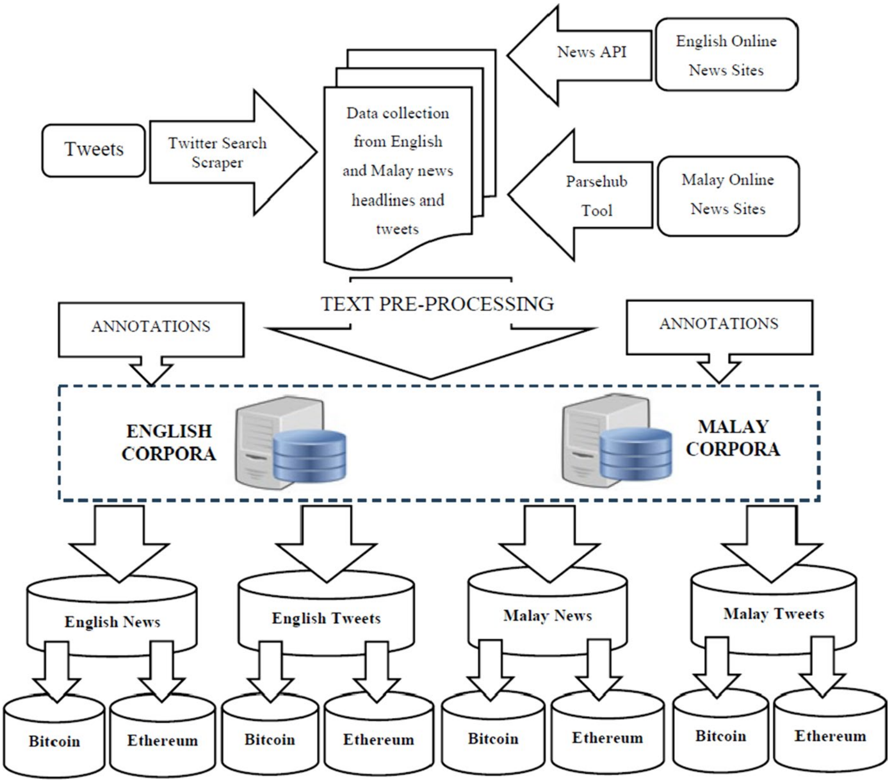


Fig. 3 Corpora Creation Framework (Zamani & Kamaruddin, 2023)

3.1 Data pre-processing

The pre-processing was performed on each of the corpus in which regular expression (regex) was applied to remove all the unwanted symbols, special characters, and hyperlinks to obtain a cleaner sentences for annotation. In this study, stopword removal and lemmatization or stemming were not used for language model as the context of the sentence is needed to learn the sentiment at sentence-level, but for regressors or classifiers experimentation, stopword removal and lemmatization were applied. This is because language models aim to capture semantic and contextual information from the text (Qiao et al., 2019), thus, removing stop words and lemmatizing words may lead to a loss of some of this contextual information, which can be crucial for understanding the meaning and relationships between words in a sentence. On the other hand, when using regressors or classifiers, removing stopwords reduces noise in the text data, making it easier for the classifier to focus on more informative words that differentiate between classes. Moreover, lemmatization reduces inflected words to their base or dictionary forms that helps in treating different inflections of a word as the same word, making it easier for the classifier to recognize word patterns and relationships (Sarica & Luo, 2021).

The experiment using language model is to generate and understand the natural language text by learning the patterns and relationships between words in a sentence by taking the text as the input to produce contextually relevant text as the output based on the language patterns learned. In this study, the predicted sentiment score will be the output after the learning of the language patterns and weights. Meanwhile, the regressors or classifiers is used to label texts into classes, in which the text is taken as the input to produce a label as the output by learning the words containing in a text instead of looking at the relationship or context between the words in a text as done in the language model.

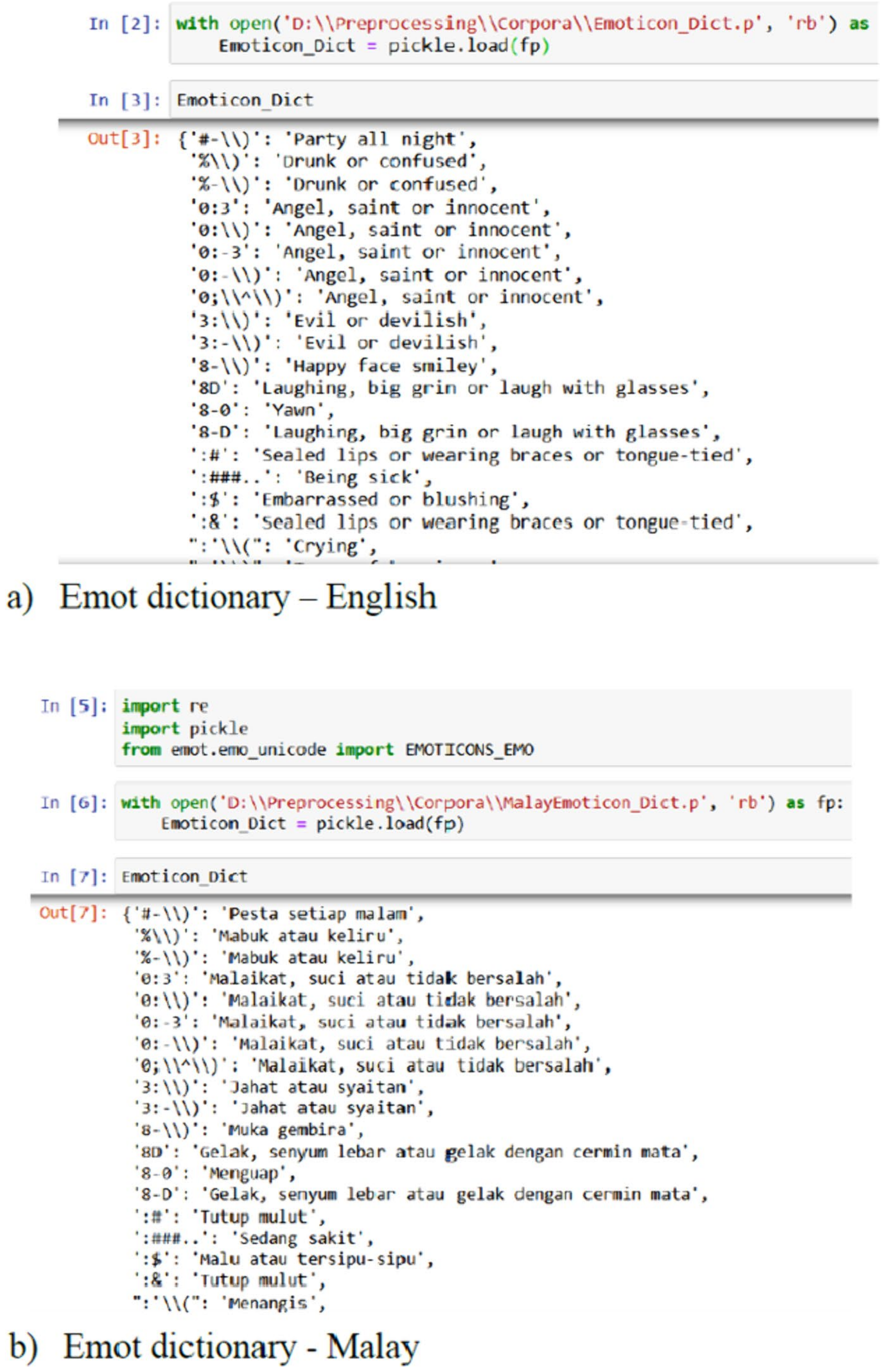
Then, the emoticons containing in some of the tweets were also handled by using Emot (Shah & Rohilla, 2022). Emot was exclusively built to translate the emoticons into English text that expresses the particular emotion, thus, in this study, the same Emot dictionary was translated into Malay text (Malay-Emot) to cater the emoticons containing in the Malay tweets corpora. The Emot dictionary for English and the Malay-Emot are displayed in Fig. 4a, b respectively.

By using the Emot dictionary, some examples of its application on the tweets containing emoticons are shown in Fig. 5a, b for English and Malay tweets correspondingly.

The cleaned corpora were then be distributed to the annotators for further process of data annotation.

3.2 Data annotation

VADER's compound score range represents the sentiment intensity lexicon that combines quantitative and qualitative features (Hutto & Gilbert, 2015), while the scoring methods by Cambria et al. (2020) used continuous labelling of range



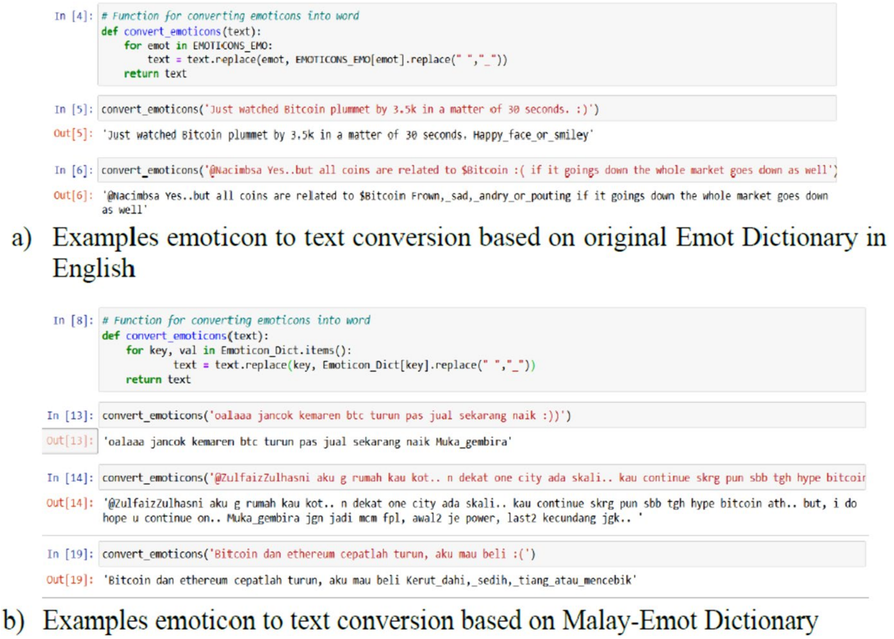
```
In [2]: with open('D:\\Preprocessing\\Corpora\\Emoticon_Dict.p', 'rb') as
            Emoticon_Dict = pickle.load(fp)

In [3]: Emoticon_Dict

Out[3]: {'#-\\)': 'Party all night',
          '%\\)': 'Drunk or confused',
          '%-\\)': 'Drunk or confused',
          '0:3': 'Angel, saint or innocent',
          '0:\\)': 'Angel, saint or innocent',
          '0:-3': 'Angel, saint or innocent',
          '0:-\\)': 'Angel, saint or innocent',
          '0;\\^\\)': 'Angel, saint or innocent',
          '3:\\)': 'Evil or devilish',
          '3:-\\)': 'Evil or devilish',
          '8-\\)': 'Happy face smiley',
          '8D': 'Laughing, big grin or laugh with glasses',
          '8-0': 'Yawn',
          '8-D': 'Laughing, big grin or laugh with glasses',
          ':#': 'Sealed lips or wearing braces or tongue-tied',
          ':###..': 'Being sick',
          ':$': 'Embarrassed or blushing',
          ':&': 'Sealed lips or wearing braces or tongue-tied',
          ":'\\(": 'Crying',
```

a) Emot dictionary – English

```
In [5]: import re
        import pickle
        from emot.emo_unicode import EMOTICONS_EMO

In [6]: with open('D:\\Preprocessing\\Corpora\\MalayEmoticon_Dict.p', 'rb') as fp:
            Emoticon_Dict = pickle.load(fp)

In [7]: Emoticon_Dict

Out[7]: {'#-\\)': 'Pesta setiap malam',
          '%\\)': 'Mabuk atau keliru',
          '%-\\)': 'Mabuk atau keliru',
          '0:3': 'Malaikat, suci atau tidak bersalah',
          '0:\\)': 'Malaikat, suci atau tidak bersalah',
          '0:-3': 'Malaikat, suci atau tidak bersalah',
          '0:-\\)': 'Malaikat, suci atau tidak bersalah',
          '0;\\^\\)': 'Malaikat, suci atau tidak bersalah',
          '3:\\)': 'Jahat atau syaitan',
          '3:-\\)': 'Jahat atau syaitan',
          '8-\\)': 'Muka gembira',
          '8D': 'Gelak, senyum lebar atau gelak dengan cermin mata',
          '8-0': 'Menguap',
          '8-D': 'Gelak, senyum lebar atau gelak dengan cermin mata',
          ':#': 'Tutup mulut',
          ':###..': 'Sedang sakit',
          ':$': 'Malu atau tersipu-sipu',
          ':&': 'Tutup mulut',
          ":'\\(": 'Menangis',
```

b) Emot dictionary - Malay

Fig. 4 Emot Dictionary



```
In [4]: # Function for converting emoticons into word
        def convert_emoticons(text):
            for emot in EMOTICONS_EMO:
                text = text.replace(emot, EMOTICONS_EMO[emot].replace(" ","_"))
            return text

In [5]: convert_emoticons('Just watched Bitcoin plummet by 3.5k in a matter of 30 seconds. :)')

Out[5]: 'Just watched Bitcoin plummet by 3.5k in a matter of 30 seconds. Happy_face_or_smiley'

In [6]: convert_emoticons('@Nacimbsa Yes..but all coins are related to $Bitcoin :( if it goings down the whole market goes down as well')

Out[6]: '@Nacimbsa Yes..but all coins are related to $Bitcoin Frown,_sad,_andry_or_pouting if it goings down the whole market goes down
        as well'
```

a) Examples emoticon to text conversion based on original Emot Dictionary in English

```
In [8]: # Function for converting emoticons into word
        def convert_emoticons(text):
            for key, val in Emoticon_Dict.items():
                    text = text.replace(key, Emoticon_Dict[key].replace(" ","_"))
            return text

In [13]: convert_emoticons('oalaaa jancok kemaren btc turun pas jual sekarang naik :))')

Out[13]: 'oalaaa jancok kemaren btc turun pas jual sekarang naik Muka_gembira'

In [14]: convert_emoticons('@ZulfaizZulhasni aku g rumah kau kot.. n dekat one city ada skali.. kau continue skrg pun sbb tgh hype bitcoir

Out[14]: '@ZulfaizZulhasni aku g rumah kau kot.. n dekat one city ada skali.. kau continue skrg pun sbb tgh hype bitcoin ath.. but, i do
         hope u continue on.. Muka_gembira jgn jadi mcm fpl, awal2 je power, last2 kecundang jgk.. '

In [19]: convert_emoticons('Bitcoin dan ethereum cepatlah turun, aku mau beli :(')

Out[19]: 'Bitcoin dan ethereum cepatlah turun, aku mau beli Kerut_dahi,_sedih,_tiang_atau_mencebik'
```

b) Examples emoticon to text conversion based on Malay-Emot Dictionary

Fig. 5 Handling emoticons in English and Malay tweets

between − 1 and + 1 (D'Orazio et al., 2022). VADER was validated using several processes in Hutto and Gilbert (2015) such as creation of a gold standard corpus, comparison with existing sentiment analysis tools, analysis of social media data, and using rule-based approach refinement. Similar process was also conducted in Cambria et al. (2020) for validity. Therefore, this study adopted this range of scores and validation methods for our corpora annotation. A rubric was designed to provide a structured and standardized set of guidelines for annotators to follow. This helps ensure that different annotators provide consistent and uniform annotations, which is critical for data quality and reliability. Table 3 shows the sentiment annotation rubrics as the sentiment scoring guidance for annotators in this study, where the total score of the rubrics will determine the annotators agreement level towards the scoring of each document.

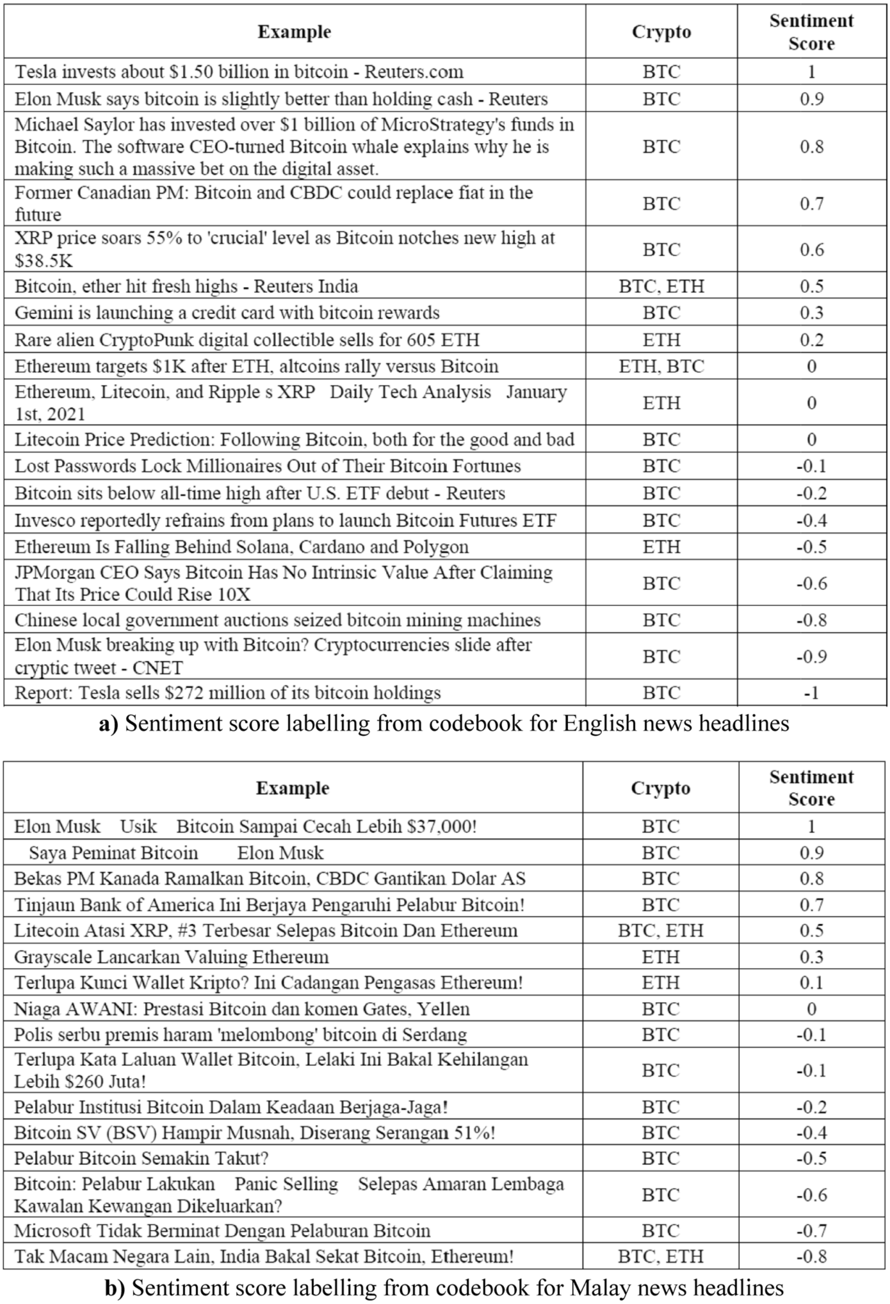
From the rubrics to obtain the standard agreement between annotators in determining the sentiment scores for each document, a codebook was then given to the annotators during the annotation tasks, where each document (i.e., a news headline or a tweet post) will be annotated with sentiment score range between − 1 and + 1, representing from very negative to very positive sentiment. A codebook provides a standardized set of guidelines and definitions for annotators to follow and crucial for assessing inter-annotator agreement among different annotators (Ritchie et al., 2022). The sentiment score focused on the issue regarding BTC or ETH (depending on the corpus) and some sample of sentiment scoring is displayed in Fig. 6a, b for English and Malay news respectively, followed by Fig. 7a, b for English and Malay tweets accordingly.

Table 3 Sentiment annotation rubrics

Categories	Very strong sentiment (3)	Strong sentiment (2)	Partial or little sentiment (1)	Confusing or no sentiment (0)
Influencer level—global/local influencer	The sentence consists of a billionnaire's name or a president of a country	The sentence consists of well-known celebrity's name or company	The sentence only stated a country's statement or announcement towards a cryptocurrency issue	The sentence is vaguely stating any influencer, or celebrity's statement towards a cryptocurrency issue
Investment action	The sentence mentions the act of highly influential people in buying or selling their cryptocurrency asset	The sentence mentions the act of well-known celebrity or company in buying or selling their cryptocurrency asset	The sentence stated a cryptocurrency user (the public) has or going to buy or sell their cryptocurrency asset	The sentence only stating general action of an individual who neither buy nor sell their cryptocurrency asset; or an individual who bought Bitcoin but sold Ethereum or vice versa for instance
Bank involvement	The sentence mentions well-known bank being involved in a very serious cryptocurrency issue; or any additional involvements between the bank and highly influential people regarding cryptocurrency matters	The sentence stated an announcement by a country's bank that can affect the behaviour of the cryptocurrency prices	The sentence consists of an individual who deals with some bank or financial institutions regarding cryptocurrency transactions	The sentence only stated a general facts about the banks involvement in cryptocurrency trading that does not have any sentiment in the context; or having mixture of positive and negative sentiment with the same weightage in the whole sentence
Personal user's opinion/ feelings	The sentence highlights a very strong emotions of a highly influential people's opinion or feelings towards a cryptocurrency	The sentence consists of the opinion or feeling from a vast amount of people in a country that leads to a serious cryptocurrency issue	The sentence only mentions an individual who was involved or just giving an opinion or feelings on a cryptocurrency either in a positive or negative perspective	The sentence did not contains any feeling nor opinion towards a cryptocurrency asset; or the sentence contains both positive and negative feelings that makes the whole sentence becomes a neutral statement

Table 3 (continued)

Categories	Very strong sentiment (3)	Strong sentiment (2)	Partial or little sentiment (1)	Confusing or no sentiment (0)
Planning or launching of cryptocurrency events	The sentence highlights a plan or launching of important cryptocurrency events by a highly influential people	The sentence consists of the announcement by a celebrity or company who plans or launching an event on cryptocurrency; or an adoption/rejection of cryptocurrency system in a country	The sentence consists of the public's events on cryptocurrency for the benefits of traders and investors in general	The sentence is only about the daily financial report that is not significant as it does not contains any sentiment nor issues that can potentially affect the cryptocurrency price movements



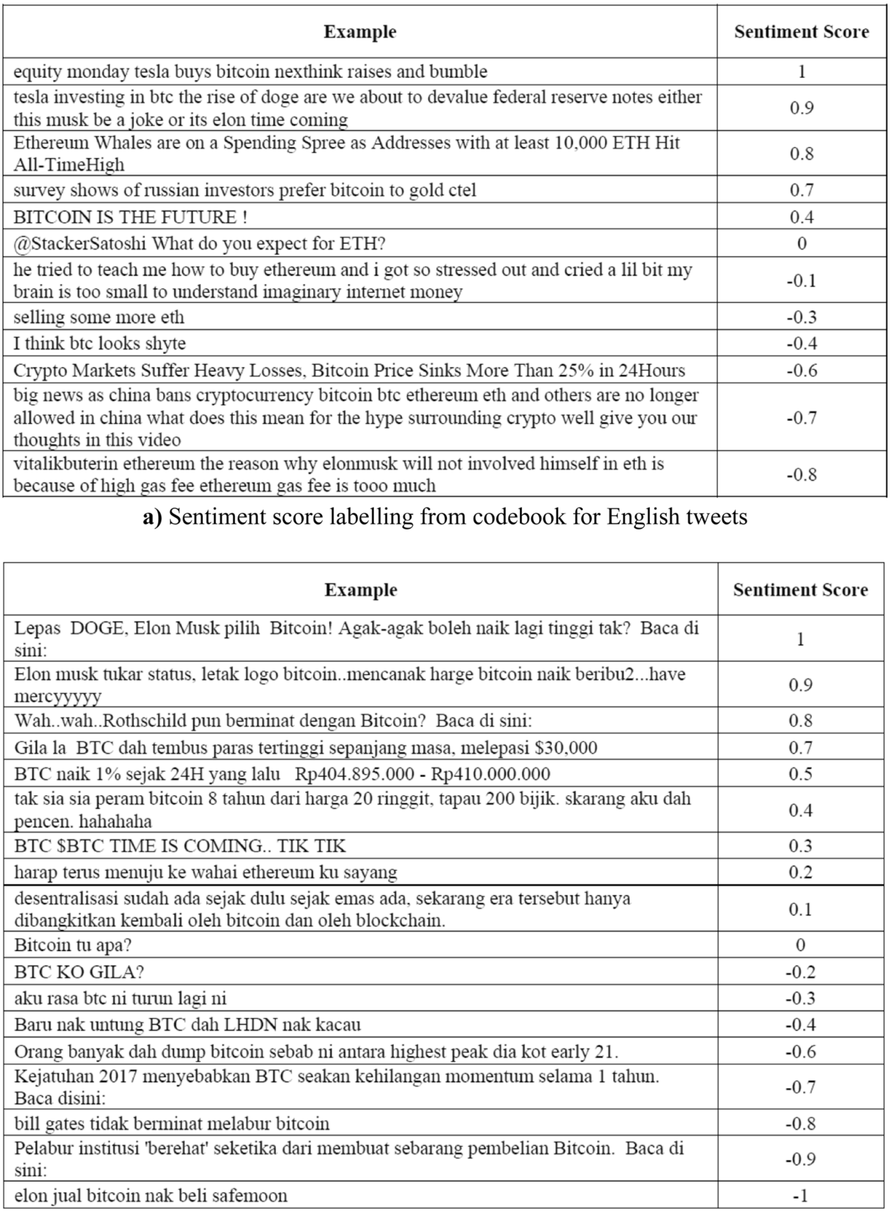
Example	Crypto	Sentiment Score
Tesla invests about $1.50 billion in bitcoin - Reuters.com	BTC	1
Elon Musk says bitcoin is slightly better than holding cash - Reuters	BTC	0.9
Michael Saylor has invested over $1 billion of MicroStrategy's funds in Bitcoin. The software CEO-turned Bitcoin whale explains why he is making such a massive bet on the digital asset.	BTC	0.8
Former Canadian PM: Bitcoin and CBDC could replace fiat in the future	BTC	0.7
XRP price soars 55% to 'crucial' level as Bitcoin notches new high at $38.5K	BTC	0.6
Bitcoin, ether hit fresh highs - Reuters India	BTC, ETH	0.5
Gemini is launching a credit card with bitcoin rewards	BTC	0.3
Rare alien CryptoPunk digital collectible sells for 605 ETH	ETH	0.2
Ethereum targets $1K after ETH, altcoins rally versus Bitcoin	ETH, BTC	0
Ethereum, Litecoin, and Ripple s XRP Daily Tech Analysis January 1st, 2021	ETH	0
Litecoin Price Prediction: Following Bitcoin, both for the good and bad	BTC	0
Lost Passwords Lock Millionaires Out of Their Bitcoin Fortunes	BTC	-0.1
Bitcoin sits below all-time high after U.S. ETF debut - Reuters	BTC	-0.2
Invesco reportedly refrains from plans to launch Bitcoin Futures ETF	BTC	-0.4
Ethereum Is Falling Behind Solana, Cardano and Polygon	ETH	-0.5
JPMorgan CEO Says Bitcoin Has No Intrinsic Value After Claiming That Its Price Could Rise 10X	BTC	-0.6
Chinese local government auctions seized bitcoin mining machines	BTC	-0.8
Elon Musk breaking up with Bitcoin? Cryptocurrencies slide after cryptic tweet - CNET	BTC	-0.9
Report: Tesla sells $272 million of its bitcoin holdings	BTC	-1

a) Sentiment score labelling from codebook for English news headlines

Example	Crypto	Sentiment Score
Elon Musk Usik Bitcoin Sampai Cecah Lebih $37,000!	BTC	1
Saya Peminat Bitcoin Elon Musk	BTC	0.9
Bekas PM Kanada Ramalkan Bitcoin, CBDC Gantikan Dolar AS	BTC	0.8
Tinjaun Bank of America Ini Berjaya Pengaruhi Pelabur Bitcoin!	BTC	0.7
Litecoin Atasi XRP, #3 Terbesar Selepas Bitcoin Dan Ethereum	BTC, ETH	0.5
Grayscale Lancarkan Valuing Ethereum	ETH	0.3
Terlupa Kunci Wallet Kripto? Ini Cadangan Pengasas Ethereum!	ETH	0.1
Niaga AWANI: Prestasi Bitcoin dan komen Gates, Yellen	BTC	0
Polis serbu premis haram 'melombong' bitcoin di Serdang	BTC	-0.1
Terlupa Kata Laluan Wallet Bitcoin, Lelaki Ini Bakal Kehilangan Lebih $260 Juta!	BTC	-0.1
Pelabur Institusi Bitcoin Dalam Keadaan Berjaga-Jaga!	BTC	-0.2
Bitcoin SV (BSV) Hampir Musnah, Diserang Serangan 51%!	BTC	-0.4
Pelabur Bitcoin Semakin Takut?	BTC	-0.5
Bitcoin: Pelabur Lakukan Panic Selling Selepas Amaran Lembaga Kawalan Kewangan Dikeluarkan?	BTC	-0.6
Microsoft Tidak Berminat Dengan Pelaburan Bitcoin	BTC	-0.7
Tak Macam Negara Lain, India Bakal Sekat Bitcoin, Ethereum!	BTC, ETH	-0.8

b) Sentiment score labelling from codebook for Malay news headlines

Fig. 6 Sentiment score labelling from codebook for news headlines



Example	Sentiment Score
equity monday tesla buys bitcoin nexthink raises and bumble	1
tesla investing in btc the rise of doge are we about to devalue federal reserve notes either this musk be a joke or its elon time coming	0.9
Ethereum Whales are on a Spending Spree as Addresses with at least 10,000 ETH Hit All-TimeHigh	0.8
survey shows of russian investors prefer bitcoin to gold ctel	0.7
BITCOIN IS THE FUTURE !	0.4
@StackerSatoshi What do you expect for ETH?	0
he tried to teach me how to buy ethereum and i got so stressed out and cried a lil bit my brain is too small to understand imaginary internet money	-0.1
selling some more eth	-0.3
I think btc looks shyte	-0.4
Crypto Markets Suffer Heavy Losses, Bitcoin Price Sinks More Than 25% in 24Hours	-0.6
big news as china bans cryptocurrency bitcoin btc ethereum eth and others are no longer allowed in china what does this mean for the hype surrounding crypto well give you our thoughts in this video	-0.7
vitalikbuterin ethereum the reason why elonmusk will not involved himself in eth is because of high gas fee ethereum gas fee is tooo much	-0.8

a) Sentiment score labelling from codebook for English tweets

Example	Sentiment Score
Lepas DOGE, Elon Musk pilih Bitcoin! Agak-agak boleh naik lagi tinggi tak? Baca di sini:	1
Elon musk tukar status, letak logo bitcoin..mencanak harge bitcoin naik beribu2...have mercyyyyy	0.9
Wah..wah..Rothschild pun berminat dengan Bitcoin? Baca di sini:	0.8
Gila la BTC dah tembus paras tertinggi sepanjang masa, melepasi $30,000	0.7
BTC naik 1% sejak 24H yang lalu Rp404.895.000 - Rp410.000.000	0.5
tak sia sia peram bitcoin 8 tahun dari harga 20 ringgit, tapau 200 bijik. skarang aku dah pencen. hahahaha	0.4
BTC $BTC TIME IS COMING.. TIK TIK	0.3
harap terus menuju ke wahai ethereum ku sayang	0.2
desentralisasi sudah ada sejak dulu sejak emas ada, sekarang era tersebut hanya dibangkitkan kembali oleh bitcoin dan oleh blockchain.	0.1
Bitcoin tu apa?	0
BTC KO GILA?	-0.2
aku rasa btc ni turun lagi ni	-0.3
Baru nak untung BTC dah LHDN nak kacau	-0.4
Orang banyak dah dump bitcoin sebab ni antara highest peak dia kot early 21.	-0.6
Kejatuhan 2017 menyebabkan BTC seakan kehilangan momentum selama 1 tahun. Baca disini:	-0.7
bill gates tidak berminat melabur bitcoin	-0.8
Pelabur institusi 'berehat' seketika dari membuat sebarang pembelian Bitcoin. Baca di sini:	-0.9
elon jual bitcoin nak beli safemoon	-1

b) Sentiment score labelling from codebook for Malay tweets

Fig. 7 Sentiment score labelling from codebook for tweets

The idea in preparing the codebook was based on relevant guides for manual annotation and defining the validity of the codebook for annotation tasks from Krippendorff (2004), Mohammad (2016), and Van Atteveldt et al. (2021). Three qualified annotators were employed to perform the manual annotation based on the designated codebook with four rounds of training and discussion on obtaining the final sentiment score for each document.

The three annotators are well-versed in English and Malay language. The first annotator is a PhD holder and a senior lecturer of language studies at Universiti Teknologi MARA with five years working experience. The first annotator annotates the data based on language knowledge, in which the labelling of each document is based on the expert understanding on the sentiment value containing in the sentence in general. The second annotator on the other hand has been involved in cryptocurrency activities for about five years and currently working as a P2P dispute specialist for almost two years who handles the cryptocurrency transactions matter between users. The third annotator is a cryptocurrency and forex trader who has been involved in these markets for about five years. The second and third annotators labelled each document based on the expert knowledge on cryptocurrency and other financial terms that imply the positivity and negativity of the statements. Positive sentiment is when the statement referring to the rising of the price and vice versa for the negative sentiment label.

The annotators were trained on standard dataset that are already labelled in prior research studies on financial-related news and social media datasets. For English news and tweets, the standard labelled dataset in continuous values was taken from John and Vechtomova (2017a), whereas for Malay news and tweets, the standard labelled datasets were obtained from Malaya Documentation by Zolkepli (2022). Krippendorf's alpha (Krippendorff, 2018) is used to evaluate the inter-annotator agreement. The Krippendorf's alpha for three annotators is shown in Eq. (1).

$$\alpha = 1 - \frac{\sum_i \sum_j \sum_k error\left(O_{ijk}\right)}{\sum_i \sum_j \sum_k error\left(E_{ijk}\right)} \tag{1}$$

where i, j, and k: Indexes for the three annotators, O_{ijk}: Observed agreement between coders i, j, and k for a specific item (unit of analysis), E_{ijk}: Expected agreement between coders i, j, and k for the same item, Error function: The error function is typically 0 when O_{ijk} equals E_{ijk} and 1 otherwise.

Once the agreement percentage reached above 60%, the training process signifies an acceptable agreement among annotators due to the subjectivity and complexity of the labelling task with difference of about ± 0.2 in allocating the sentiment scores (one decimal place). Hence, this study has obtained the average percentage agreement of 65% and 63% for English and Malay news headlines respectively, while the average percentage agreement of 62% and 60% were obtained for English and Malay tweets accordingly. Then, based on the agreement, this study extended the corpora to produce the eight corpora using a more cryptocurrency-specific texts focusing on Bitcoin and Ethereum matters. The inter-annotator agreement was validated by another expert in the financial market, who has a 25-years working experience as a remisier at Kenanga Investment Bank, Malaysia, dealing with various financial markets such as stock market, forex, and cryptocurrency, and was not directly involved in the annotation process.

4 Experimental setup

This study implemented a sentiment model by using Generalized Autoregressive Pretraining for Language Understanding (XLNet) language model and Bidirectional Gated Recurrenct Unit (Bi-GRU) with muti-head self-attention to evaluate the performance of the annotated dataset. This is because language modelling application has shown a good result in sentiment analysis as reported in Passalis et al. (2022) and Cerda (2021) that applied BERT language model. BERT functions by using fixed forward–backward factorisation order, while XLNet was introduced after BERT that functions by considering all possible permutations of the factorisation order and also does not undergo pretrain-finetune discrepancy as BERT does (Yang et al., 2020). In addition to the XLNet language model, a Bidirectional GRU deep learning model was chosen instead of the commonly used LSTM in other prior research studies since GRU has shown to perform better in dealing with smaller datasets and with the bidirectional behaviour of the GRU, it can enhance the learning of the relationships between words with forward–backward movements (Endalie et al., 2022). The multi-head self-attention mechanism introduced by Vaswani et al. (2017) will also be integrated into the GRU layer with the purpose to focus on the learning of semantic information in the sentences by capturing the significant words associated with sentiment towards the cryptocurrency issue (Li et al., 2022) and also assists the model from overfitting (Zhang et al., 2023). Hence, the train-test split for the experiment is 80% train set and 20% test set for each corpus. Figure 8 illustrates the sentiment model framework.

Based on Fig. 8, each sentence from the corpus is taken as the input to be tokenized with XLNetTokenizer ("xlnet-base-cased" for English text and "xlnet-base-bahasa-cased" for Malay text) consisting of 12 layers of transformer blocks, 768 hidden layers (dimensions), and 12 self-attention heads (Gong et al., 2019). Next, the encoding of data was performed and produced context vector to proceed with the fine-tuning of pre-trained XLNet language model. The fine-tuning was done to learn the new weights from the annotated corpora. The fine-tuned XLNet was then fed into the Bi-GRU layer which consists of 16 hidden dimensions in one layer, and a dropout value of 0.5 with one output value using Tanh activation function. Tanh was applied since it caters the range of output between − 1 and 1 aligning with the scope of this study. AdamW optimizer was also applied with training learning rate of 2e−5. A fivefold cross validation of GridSearch was used to find the most optimised hyperparameter settings. Multi-head self-attention mechanism by Vaswani et al. (2017) was inserted into the implementation to gain a better learning quality of the sentiment model.

The annotated corpora is also experimented with other regressors applied in prior existing research studies to compare the performance achieved using the datasets to avoid bias in the evaluation. Since there are very limited study on Malay language, experiments conducted on English texts were used as the

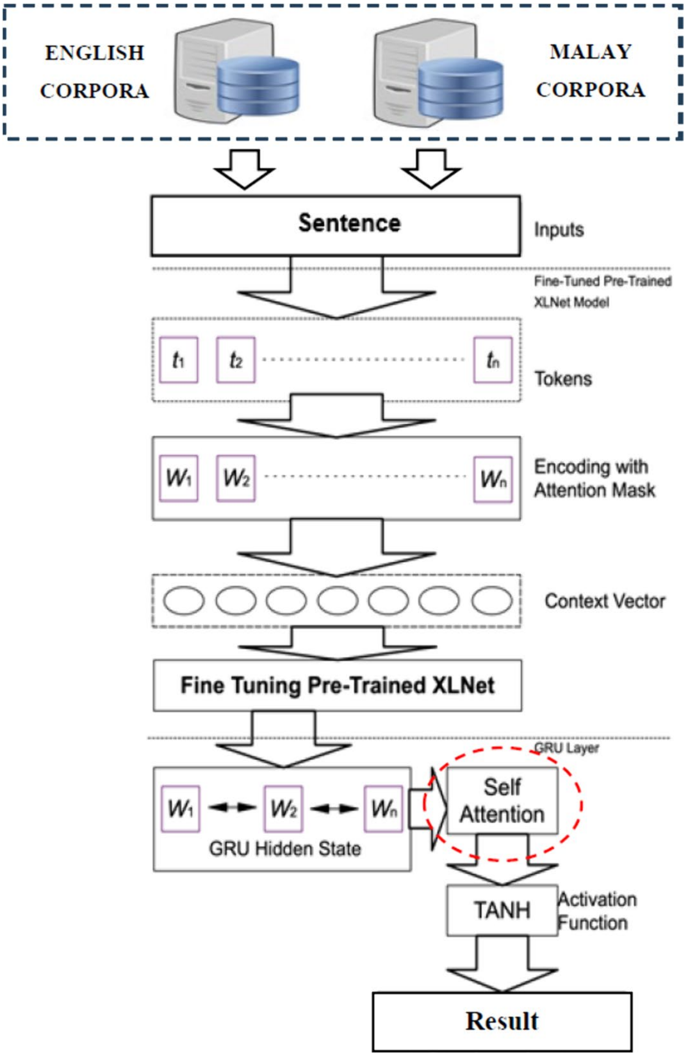


Fig. 8 XLNet-BiGRU-MHA model framework

benchmark albeit the results might be slightly lower due to the very low language resources available for Malay.

Several regression models will be explored with the XLNet-BiGRU-MHA sentiment model, which are Linear Regression (John & Vechtomova, 2017b; Symeonidis et al., 2017), Support Vector Regression (John & Vechtomova, 2017b), Random Forest, and Decision Tree (Symeonidis et al., 2017). In addition to the regression experiments, BERT language model with GRU deep learning model was also applied to examine the performance of another language model that was commonly used in several related works in using the annotated dataset with continuous labels.

Three evaluation metrics are used which are Mean Absolute Error (MAE), Root Mean Square Error (RMSE), R^2, and adjusted R^2 for regression model comparisons, while accuracy is applied to investigate the performance on classification models.

5 Results and discussion

5.1 Experiments on dataset using regressors

Table 4 presents the evaluation results achieved for news corpora, while and Table 5 shows the evaluation results obtained for tweets corpus.

Table 4 Evaluation on regression models (news headlines)

Corpus	Cryptocurrency	Model	MAE	RMSE	R^2	Adj R^2	Accuracy (%)
English news	Bitcoin	Linear Regression/ Logistic Regression	0.274	0.378	0.439	0.436	70
		SVM	0.303	0.388	0.408	0.405	75.5
		Random Forest	0.305	0.432	0.267	0.264	70.5
		Decision Tree	0.322	0.497	0.029	0.024	69
		BERT-GRU	0.180	0.334	0.562	0.559	84.5
		XLNet-BiGRU-MHA	**0.165**	**0.265**	**0.724**	**0.722**	**86**
	Ethereum	Linear Regression/ Logistic Regression	0.198	0.293	0.347	0.344	72.5
		SVM	0.250	0.308	0.279	0.276	76.9
		Random Forest	0.236	0.332	0.165	0.161	70
		Decision Tree	0.242	0.398	− 0.204	− 0.210	64.5
		BERT-GRU	0.163	0.284	0.388	0.384	77.5
		XLNet-BiGRU-MHA	**0.074**	**0.156**	**0.816**	**0.815**	**79**
Malay news	Bitcoin	Linear Regression/ Logistic Regression	0.287	0.373	0.365	0.361	80
		SVM	0.314	0.377	0.351	0.348	80
		Random Forest	0.264	0.372	0.369	0.365	79
		Decision Tree	0.292	0.463	0.022	0.017	77
		BERT-GRU	0.211	0.350	0.450	0.440	85.7
		XLNet-BiGRU-MHA	**0.219**	**0.327**	**0.514**	**0.513**	**91.5**
	Ethereum	Linear Regression/ Logistic Regression	0.227	0.316	0.456	0.453	84.5
		SVM	0.253	0.316	0.457	0.454	85.5
		Random Forest	0.197	0.307	0.487	0.484	85.1
		Decision Tree	0.218	0.390	0.173	0.169	79
		BERT-GRU	0.208	0.274	0.600	0.589	87
		XLNet-BiGRU-MHA	**0.141**	**0.251**	**0.656**	**0.654**	**90**

Table 5 Evaluation on regression models (tweets)

Corpus	Cryptocurrency	Model	MAE	RMSE	R^2	Adj R^2	Accuracy (%)
English tweets	Bitcoin	Linear Regression/Logistic Regression	0.333	0.453	−1.414	− 1.416	64.4
		SVM	0.200	0.262	0.191	0.191	65.1
		Random Forest	0.200	0.272	0.128	0.127	62.2
		Decision Tree	0.236	0.342	− 0.376	− 0.377	53.5
		BERT-GRU	0.148	0.221	0.427	0.427	68.3
		XLNet-BiGRU-MHA	**0.139**	**0.213**	**0.465**	**0.464**	**92.7**
	Ethereum	Linear Regression/Logistic Regression	0.269	0.356	− 1.353	− 1.355	67.6
		SVM	0.168	0.216	0.143	0.143	68.4
		Random Forest	0.159	0.226	0.063	0.063	66.9
		Decision Tree	0.191	0.288	− 0.521	− 0.522	59.3
		BERT-GRU	0.120	0.188	0.351	0.350	71.5
		XLNet-BiGRU-MHA	**0.122**	**0.184**	**0.380**	**0.379**	**97.3**
Malay tweets	Bitcoin	Linear Regression/Logistic Regression	0.346	0.470	− 0.525	− 0.526	76.6
		SVM	0.228	0.288	0.430	0.430	76.2
		Random Forest	0.200	0.290	0.421	0.420	74.25
		Decision Tree	0.238	0.373	0.042	0.042	70.1
		BERT-GRU	0.208	0.288	0.430	0.428	78.3
		XLNet-BiGRU-MHA	**0.183**	**0.281**	**0.457**	**0.457**	**88**
	Ethereum	Linear Regression/Logistic Regression	0.293	0.400	− 0.752	− 0.753	77.5
		SVM	0.194	0.241	0.353	0.353	76.25
		Random Forest	0.178	0.255	0.275	0.275	70.5
		Decision Tree	0.210	0.320	− 0.143	− 0.144	71
		BERT-GRU	0.173	0.242	0.350	0.349	77.8
		XLNet-BiGRU-MHA	**0.157**	**0.233**	**0.400**	**0.393**	**87**

From Table 4, it can be observed that our XLNet-BiGRU-MHA model performs the best throughout all four news corpora. It is found that sentiment regression on Ethereum for English and Malay news achieved better result than Bitcoin by about ± 0.1 in the adjusted R^2. This may be due to the lesser number of documents during training for Ethereum since the public's interest is more towards Bitcoin. However, for other regressors, a larger sized dataset will be more significant to improve the outcomes. It can be seen that applying language model as text representation for the deep learning model boosts the precision of the sentiment prediction. Furthermore, an adjusted R^2 of − 0.210 was obtained by the decision tree for English news on Ethereum, which denotes a poor fit in accommodating sentiment words. The rationale of this is due to the inconsideration of context in the sentences when lexicon-based approach such as TF-IDF or Bag-of-Words is applied. On the contrary, text representations techniques such as word embeddings (e.g., Word2Vec, GloVe) and language models (e.g., BERT, XLNet) manoeuvre such algorithm that take the context of sentences into account by producing a certain word vectors.

In comparison with the results obtained in John and Vechtomova (2017a, 2017b) employing 1143 instances in the news dataset, it is reported that only an R^2 of − 0.1 was achieved using TF-IDF and Linear Regression, while an R^2 of 0.38 was obtained using TF-IDF and Support Vector Regression. This infers that our annotated corpora is a better fit for Linear Regression with an average R^2 of 0.393 for English news and an average R^2 0.411 for Malay news. This infers that our Malay corpora manage to perform well on Linear Regression model even though limited in language resources.

Mixed results was achieved for the Support Vector Regression when comparing with John and Vechtomova (2017a, 2017b) experiment. This study shows that the English news on Bitcoin and Malay news on Ethereum corpora managed to outperforms John and Vechtomova (2017a, 2017b) by 0.028 and 0.077 in the R^2 value respectively. However, there is a slightly lower R^2 obtained for our Ethereum English news and Bitcoin Malay news with an R^2 of 0.279 and 0.351 correspondingly. The results also suggest that our corpora can fit moderately using the Support Vector Regressor. Another comparisons on tweets data are shown in Table 5.

Our XLNet-BiGRU-MHA model also shows the best performed sentiment regression model in Table 5, where the number of each tweets corpus for training is the same. Nevertheless, both English and Malay tweets on Bitcoin dataset have achieved higher performance results compared to Ethereum dataset. In spite of this, the overall achievements between news and tweets corpora shows that tweets data give a lower precision than news data. This is because the texts used in tweets contain more noise and lack of proper sentence structure due to the usage of acronyms, spelling errors, mixed languages, emojis and emoticons, which leads to a more challenging text pre-processing, thus, making the learning process of the machines becomes more complicated. From Table 5, it can be observed that all four tweets corpora do not fit into linear regression model. The same outcome is also displayed for the decision tree model except for Malay tweets on Bitcoin that achieved a very low adjusted R^2 of 0.042. The same proposition in increasing the number of datasets for training can also be applied for tweets corpora. In terms of language, we can discern that English texts give a superior performance than Malay texts as Malay is deemed to be a low resource language, while English is used worldwide and has more resources to enhance the algorithms in computational tasks.

In terms of the errors, our corpora achieved an average MAE of 0.301 and RMSE of 0.405 for English tweets, while an average MAE of 0.320 and RMSE of 0.435 for Malay tweets using Linear Regression. This has also shown a better result compared to the experiment by Symeonidis et al. (2017) on 2333 tweets data using Linear Regression obtaining an MAE of 0.446 and RMSE of 0.568. Whilst, the employment of Decision Tree in Symeonidis et al. (2017) reported an MAE of 0.384 and RMSE of 0.472. Correspondingly, our dataset obtained an average MAE of 0.214 and an average RMSE of 0.315 for English tweets corpora, while Malay tweets corpora attained an average MAE of 0.224 and an average RMSE of 0.347. This shows that although the adjusted R^2 produced negative results denoting poor fit of the model with our dataset, the prediction errors still gave a slightly lesser value than Symeonidis et al. (2017) work.

Correspondingly, based on the accuracy shown in Tables 4 and 5, the XLNet-BiGRU-MHA model has shown to produce a good accuracy in evaluating classification model and our corpora has shown to perform well with the traditional regressors and classifiers.

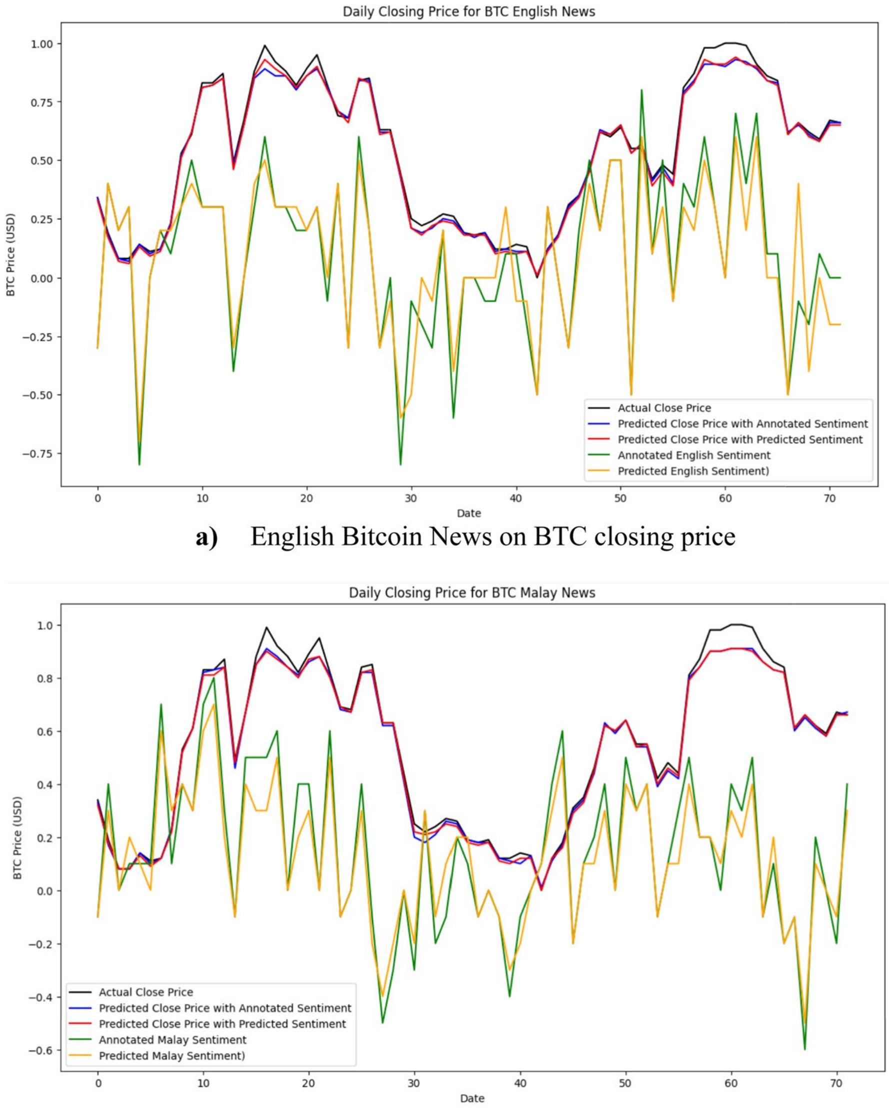
5.2 Sentiment analysis towards cryptocurrency price trends

The Bitcoin (BTC) and Ethereum (ETH) prices are predicted by aggregating five daily price features: open, high, low, close prices, and trading volume with the overall (average) daily sentiment score feature. Then, data normalization was performed for the prices. The normalized data were fed into a three-layered Bi-GRU deep learning model for data training. The train-test split for the experiment is 80–20%. The hyperparameter settings for the price prediction model is shown in Table 6, where k-fold GridSearch cross validation was used to optimised the settings.

The results of the predicted prices are shown in Figs. 9, 10, 11 and 12 displaying the five different lines in each graph. The actual closing price shown is the BTC or ETH daily close price obtained from Bitstamp exchange, while the two predicted prices shown are: (1) the predicted price obtained using the OHLC price features with added sentiment score feature through manual annotation, and (2) the predicted price obtained using the OHLC price features and automatically generated sentiment scores via our XLNet-BiGRU-MHA model. The last two

Table 6 Price Prediction Hyperparameter Setting

Setting	Value
Hidden unit	50
Dropout	0.2
Dense layer	1
Optimizer	Adam
Learning rate	0.001
Activation function	Tanh



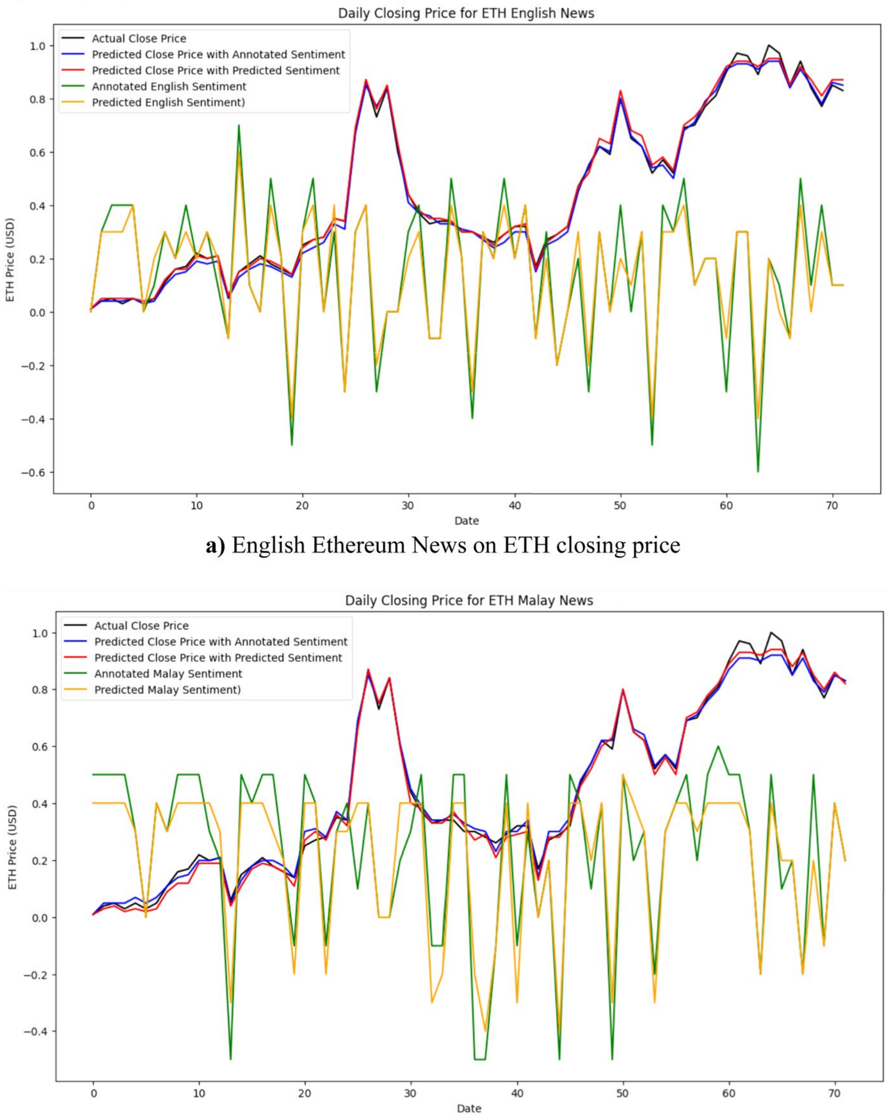
a) English Bitcoin News on BTC closing price

b) Malay Bitcoin News on BTC closing price

Fig. 9 Bitcoin news on BTC closing price

sentiment lines displayed are the manually annotated sentiment score and the sentiment score automatically generated from the XLNet-BiGRU-MHA model (daily overall sentiment score).

From Fig. 9, 10, 11 and 12, the y-axis represents the normalized prices of BTC and ETH ranging from 0 to 1 to scale proportionately to the sentiment polarity ranging from − 1 to + 1. The price range for BTC prices is from $46,210.92 to $67,559.00, and the price range for ETH prices is between $3,627.68 and $4,811.59.



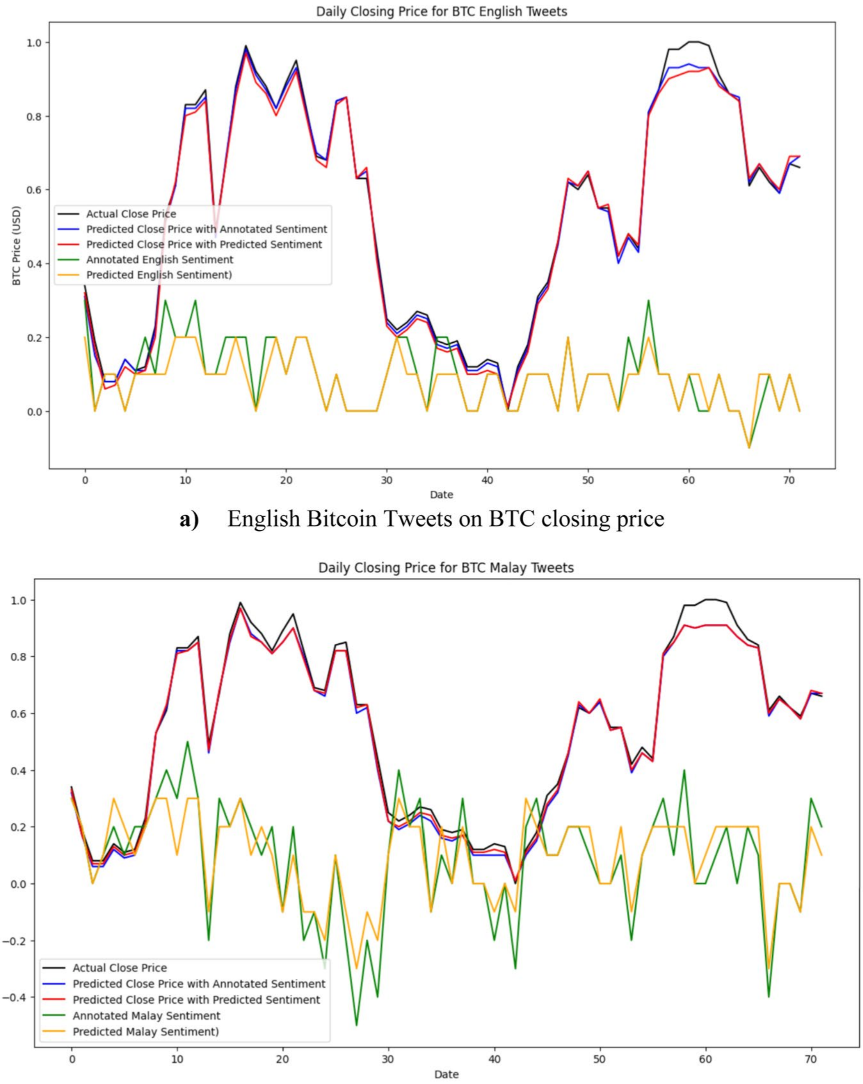
a) English Ethereum News on ETH closing price

b) Malay Ethereum News on ETH closing price

Fig. 10 Ethereum news on ETH closing price

The x-axis represents the days from the testing set that comprises of 73 days (from 20 October 2021 to 31 December 2021) out of the 365 days (from 1 January 2021 to 31 December 2021). From Figs. 9, 10, 11 and 12, it can be observed that sentiment does play a role on the cryptocurrency price directions.

Positive sentiment will result in higher BTC price whereas negative sentiment will yield downward trend for BTC price. It can be observed in Figs. 9 and



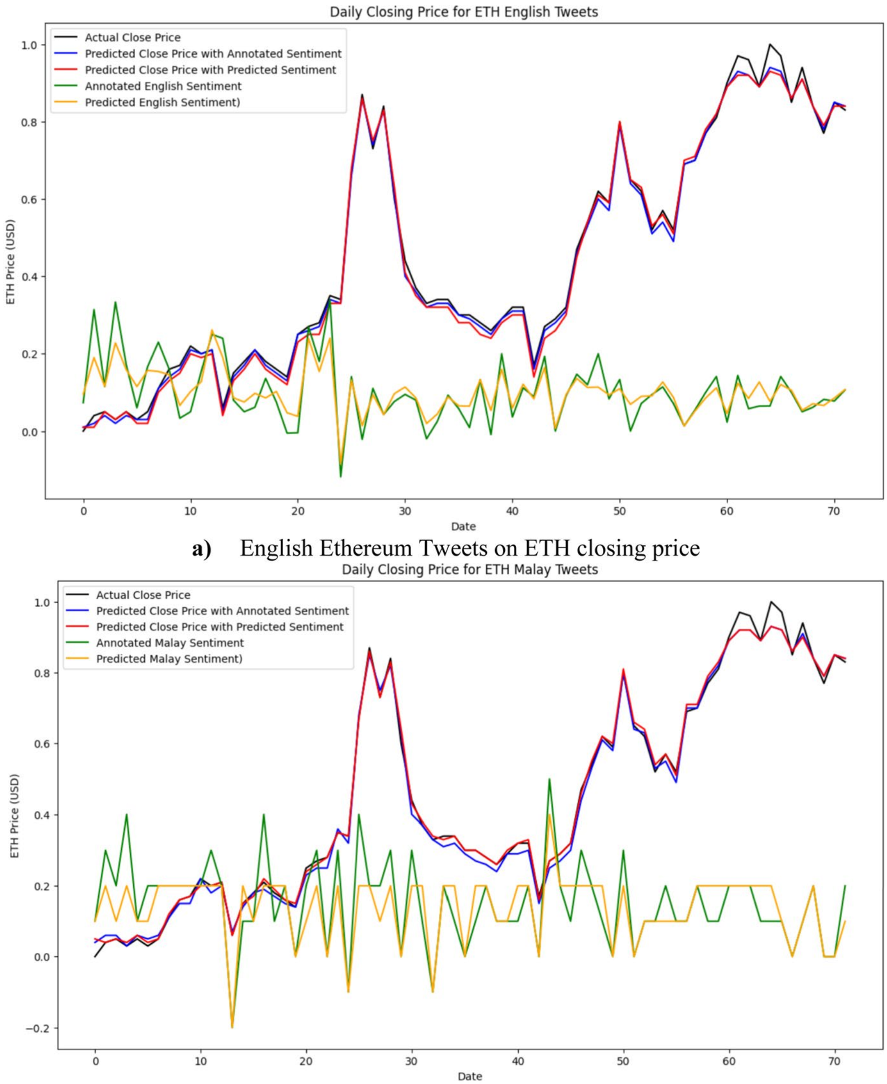
a) English Bitcoin Tweets on BTC closing price

b) Malay Bitcoin Tweets on BTC closing price

Fig. 11 Bitcoin tweets on BTC closing price

11. From Fig. 9a, it can be seen that positive sentiments from 26 October 2021 (x = 7) until 14 November 2021 (x = 26) show the increment of BTC prices. On the contrary, during the period of 20 November 2021 (x = 32) until 30 November 2021 (x = 42), the sentiments are negative which reflects the decrement of BTC prices. This pattern is consistent for all graphs (Figs. 10, 11, 12).

From Figs. 9 and 11, the increment and decrement of price pattern is more consistent because of the gradual change in the sentiment. However, the price



a) English Ethereum Tweets on ETH closing price

b) Malay Ethereum Tweets on ETH closing price

Fig. 12 Ethereum tweets on ETH closing price

pattern in Figs. 10 and 12, shows some spikes because of the sudden change in the sentiment. It implies the direct impact of sentiment towards the prices.

Hence, it can be inferred that both sentiment from tweets and news does follow the same patterns as the cryptocurrency price movements, but in terms of the extreme positive and negative sentiment patterns towards the price differs slightly between the news and tweets. News sentiment shows many drastic change in the sentiment that affecting the prices, whereas tweets sentiment has more neutral score at the time the price were slightly increase or decrease. This may be due to the number of tweets that contains more public's opinion which may not give much impact on the prices compared to the official online news being published. Online news usually highlights the important issues more than tweets by public users (Reis et al., 2015). Tweets are usually more impactful if the postings were from global influencers such as Elon Musk or announcements by certain leader of a country. Moreover, this work also aligns with the work by Wan et al. (2021) that investigates the social media effects towards the market movements.

In terms of the prediction between the actual and predicted sentiment using the XLNet-BiGRU-MHA sentiment model, it can be observed the small gaps in the prediction errors showing the reliability of the implemented sentiment model to perform the sentiment regression on cryptocurrency domain. In comparison with previous related studies, the results showing significant impact of sentiment towards the cryptocurrency price patterns is inline with Passalis et al. (2022) work that applied sentiment feature in addition to the closing price achieving an accuracy of 92% for the price prediction result. In addition, El Haddaoui et al. (2023) has also shown to achieve good price prediction result with an R^2 of 0.986 (regression task).

6 Conclusion

Based on the literature review in Sect. 2, our work contributes in creating sentiment-based cryptocurrency-specific (Bitcoin and Ethereum) news as well as tweets corpora in English and Malay languages. Instead of using automated tools for sentiment labelling, we performed manual annotation with training process among the annotators. Manual annotation allows for high-quality and accurate data labeling. The nuance of the language can be picked up in the right context especially the annotators are the expert in the field. Humans can also handle ambiguous or context-dependent data more effectively than automated methods by interpreting and disambiguate information when the context is unclear. On top of that, the sentiment agreement by the annotators are validated by the other expert who is not involve directly in the annotation process, thus it can minimize bias.

The limitation of our corpora is that only the data throughout the year 2021 was extracted for English and Malay as it was the peak of the pandemic COVID-19 season, whereby online platforms were actively utilized. Another limitation is the low resource Malay language dictionary or language model to be used for computation that may lead to less word coverage than the English language resources. From the outcomes presented in Tables 4 and 5, our corpora are practical to be experimented with various machine learning and deep learning models for sentiment analysis. Although a larger sized corpus is needed to enhance the accuracy of the typical regressors, our corpora performed well with language models utilisation. This shows that language models is significant in training lesser number of data. In comparison with BERT language model, our XLNet-BiGRU-MHA sentiment regression model achieved smaller errors and higher adjusted R^2 result. Hence, it signifies that our XLNet-BiGRU-MHA model performance is at par with the current studies.

Therefore, our corpora creates an opportunity for other researchers to perform more experiments on cryptocurrency and other financial domain. For our future endeavour, we will gather more data focusing on the current year's news and tweets to enlarge the current corpora, consequently will improve the accuracy of the machine learning and deep learning models.

Acknowledgements The authors would like to thank the annotators who participated in the creation of the Malay news and tweets corpora from Universiti Teknologi MARA (UiTM) and Universiti Sains Malaysia (USM).

Author Contributions Mohamad Zamani: Conceptualization, Methodology, Data Curation, Investigation, Writing – Original Draft. Kamaruddin: Writing – Review & Editing, Investigation. Yusof: Writing – Editing.

Funding This research received no specific grant from any funding agency.

Data availability Available upon request.

Declarations

Conflicts of interest The authors declare no conflict of interest.

References

Agarwal, B., Harjule, P., Chouhan, L., Saraswat, U., Airan, H., & Agarwal, P. (2021). Prediction of dogecoin price using deep learning and social media trends. *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems, 8*(29), 171188. https://doi.org/10.4108/eai.29-9-2021.171188

Aggarwal, A., Gupta, I., Garg, N., & Goel, A. (2019). Deep learning approach to determine the impact of socio economic factors on Bitcoin price prediction. *2019 Twelfth International Conference on Contemporary Computing (IC3)*, 1–5. https://doi.org/10.1109/IC3.2019.8844928

Ahmad, W., Wang, B., Martin, P., Xu, M., & Xu, H. (2023). Enhanced sentiment analysis regarding COVID-19 news from global channels. *Journal of Computational Social Science, 6*(1), 19–57. https://doi.org/10.1007/s42001-022-00189-1

Alghamdi, S., Alqethami, S., Alsubait, T., & Alhakami, H. (2022). Cryptocurrency price prediction using forecasting and sentiment analysis. *International Journal of Advanced Computer Science and Applications*. https://doi.org/10.14569/IJACSA.2022.01310105

Althnian, A., AlSaeed, D., Al-Baity, H., Samha, A., Dris, A. B., Alzakari, N., Abou Elwafa, A., & Kurdi, H. (2021). Impact of dataset size on classification performance: An empirical evaluation in the medical domain. *Applied Sciences, 11*(2), 796. https://doi.org/10.3390/app11020796

Attila, S. D. (2017). Impact of social media on cryptocurrency trading with deep learning. *Scientific Students' Conference* (p. 47).

Balfagih, A. M., & Keselj, V. (2019). Evaluating sentiment classifiers for Bitcoin tweets in price prediction task. *2019 IEEE International Conference on Big Data (Big Data)* (pp. 5499–5506). https://doi.org/10.1109/BigData47090.2019.9006140

Barbaglia, L., Frattarolo, L., Onorante, L., Pericoli, F. M., Ratto, M., & Tiozzo Pezzoli, L. (2022). Testing big data in a big crisis: Nowcasting under Covid-19. *International Journal of Forecasting*. https://doi.org/10.1016/j.ijforecast.2022.10.005

Bonta, V., Kumaresh, N., & Janardhan, N. (2019). A comprehensive study on lexicon based approaches for sentiment analysis. *Asian Journal of Computer Science and Technology, 8*(S2), 1–6. https://doi.org/10.51983/ajcst-2019.8.S2.2037

Cambria, E., Li, Y., Xing, F. Z., Poria, S., & Kwok, K. (2020). SenticNet 6: Ensemble Application of Symbolic and Subsymbolic AI for Sentiment Analysis. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (pp. 105–114). https://doi.org/10.1145/3340531.3412003

Cerda, G. N. C. (2021). *Bitcoin price prediction through stimulus analysis: On the footprints of Twitter's crypto-influencers* [Master's Thesis, Pontificia Universidad Católica de Chile]. https://repositorio.uc.cl/xmlui/bitstream/handle/11534/60881/TESIS_GCheuque_Firma%20Final.pdf?sequence=1

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Editorial: Special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter, 6*(1), 1–6. https://doi.org/10.1145/1007730.1007733

Chen, C. Y.-H., Després, R., Guo, L., & Renault, T. (2019). What makes cryptocurrencies special? Investor sentiment and return predictability during the bubble. *Comparative Political Economy: Monetary Policy eJournal* 1–36.

Chin, C. K., & Omar, N. (2020). Bitcoin price prediction based on sentiment of news article and market data with LSTM model. *Asia-Pacific Journal of Information Technology and Multimedia, 9*(1), 1–16. https://doi.org/10.17576/apjitm-2020-0901-01

Chowdhury, R., Rahman, M. A., Rahman, M. S., & Mahdy, M. R. C. (2020). An approach to predict and forecast the price of constituents and index of cryptocurrency using machine learning. *Physica a: Statistical Mechanics and Its Applications, 551*, 124569. https://doi.org/10.1016/j.physa.2020.124569

Corbet, S., Hou, Y., Hu, Y., Larkin, C., Lucey, B., & Oxley, L. (2022). Cryptocurrency liquidity and volatility interrelationships during the COVID-19 pandemic. *Finance Research Letters, 45*, 102137. https://doi.org/10.1016/j.frl.2021.102137

Critien, J. V., Gatt, A., & Ellul, J. (2022). Bitcoin price change and trend prediction through twitter sentiment and data volume. *Financial Innovation, 8*(1), 45. https://doi.org/10.1186/s40854-022-00352-7

Daskalakis, N., & Georgitseas, P. (2020). *An Introduction to Cryptocurrencies: The Crypto Market Ecosystem*. Routledge.

D'Orazio, M., Di Giuseppe, E., & Bernardini, G. (2022). Automatic detection of maintenance requests: Comparison of Human Manual Annotation and Sentiment Analysis techniques. *Automation in Construction, 134*, 104068. https://doi.org/10.1016/j.autcon.2021.104068

Edgari, E., Thiojaya, J., & Qomariyah, N. N. (2022). The impact of Twitter sentiment analysis on Bitcoin price during COVID-19 with XGBoost. *2022 5th International Conference on Computing and Informatics (ICCI)* (pp. 337–342). https://doi.org/10.1109/ICCI54321.2022.9756123

El Haddaoui, B., Chiheb, R., Faizi, R., & El Afia, A. (2023). The influence of social media on cryptocurrency price: A sentiment analysis approach. *International Journal of Computing and Digital Systems, 13*(1), 1–15. https://doi.org/10.12785/ijcds/130137

Endalie, D., Haile, G., & Taye, W. (2022). Bi-directional long short term memory-gated recurrent unit model for Amharic next word prediction. *PLoS ONE, 17*(8), e0273156. https://doi.org/10.1371/journal.pone.0273156

Farhana, K., & Muthaiyah, S. (2022). Behavioral intention to use cryptocurrency as an electronic payment in Malaysia. *Journal of System and management Science, 12*(4), 219–231.

Galeshchuk, S., Vasylchyshyn, O., & Krysovatyy, A. (2018). Bitcoin response to Twitter sentiments. *ICTERI Workshops* (pp. 160–168).

Garg, A., Shah, T., Jain, V. K., & Sharma, R. (2021). CrypTop12: A dataset for cryptocurrency price movement prediction from tweets and historical prices. *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 379–384). https://doi.org/10.1109/ICMLA52953.2021.00065

Goleman, T. (2018). *Cryptocurrency: Mining, investing and trading in Blockchain for Beginners. How to buy cryptocurrencies (Bitcoin, Ethereum, Ripple, Litecoin or Dash) and what wallet to use. Cryptocurrency investment strategies*. Zen Mastery.

Gong, X.-R., Jin, J.-X., & Zhang, T. (2019). Sentiment analysis using autoregressive language modeling and broad learning system. *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 1130–1134). https://doi.org/10.1109/BIBM47256.2019.8983025

Gurdgiev, C., & O'Loughlin, D. (2020). Herding and anchoring in cryptocurrency markets: Investor reaction to fear and uncertainty. *Journal of Behavioral and Experimental Finance, 25*, 100271. https://doi.org/10.1016/j.jbef.2020.100271

Hartmann, J., Heitmann, M., Siebert, C., & Schamp, C. (2023). More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing, 40*(1), 75–87. https://doi.org/10.1016/j.ijresmar.2022.05.005

Hasan, S. H., Hasan, S. H., Ahmed, M. S., & Hasan, S. H. (2022). A Novel cryptocurrency prediction method using optimum CNN. *Computers, Materials & Continua, 71*(1), 1051–1063. https://doi.org/10.32604/cmc.2022.020823

Hitam, N. A., Ismail, A. R., Samsudin, R., & Ameerbakhsh, O. (2021). The influence of sentiments in digital currency prediction using hybrid sentiment-based support vector machine with whale optimization algorithm (SVMWOA). *International Congress of Advanced Technology and Engineering (ICOTEN)*. https://doi.org/10.1109/ICOTEN52080.2021.9493454

Hooson, M., & Pratt, K. (2023, October 4). Our Pick Of The Best Cryptocurrencies Of October 2023. *Forbes Advisor*. https://www.forbes.com/uk/advisor/investing/cryptocurrency/top-10-cryptocurrencies-october-2023/

Hu, M., & Liu, B. (2004). *Mining and Summarizing Customer Reviews*.

Hutto, C. J., & Gilbert, E. (2015). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM 2014)*.

Ibrahim, A. (2021). Forecasting the early market movement in Bitcoin using Twitter's sentiment analysis: An ensemble-based prediction model. *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)* (pp. 1–5). https://doi.org/10.1109/IEMTRONICS52119.2021.9422647

Inamdar, A., Bhagtani, A., Bhatt, S., & Shetty, P. M. (2019). Predicting cryptocurrency value using sentiment analysis. *2019 International Conference on Intelligent Computing and Control Systems (ICCS)* (pp. 932–934). https://doi.org/10.1109/ICCS45141.2019.9065838

Jahjah, F. H., & Rajab, M. (2020). Impact of Twitter sentiment related to bitcoin on stock price returns. *Journal of Engineering, 26*(6), 60–71. https://doi.org/10.31026/j.eng.2020.06.05

Jain, A., Tripathi, S., Dwivedi, H. D., & Saxena, P. (2018). Forecasting price of cryptocurrencies using tweets sentiment analysis. *2018 Eleventh International Conference on Contemporary Computing (IC3)* (pp. 1–7). https://doi.org/10.1109/IC3.2018.8530659

John, V., & Vechtomova, O. (2017a). *SemEval-2017 Task 5 News and Microblogs dataset*. SemEval-2017 Task 5. https://alt.qcri.org/semeval2017/task5/index.php?id=data-and-tools

John, V., & Vechtomova, O. (2017b). UW-FinSent at SemEval-2017 Task 5: Sentiment analysis on financial news headlines using training dataset augmentation. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 872–876). https://doi.org/10.18653/v1/S17-2149

Kang, C. Y., Lee, C. P., & Lim, K. M. (2022). Cryptocurrency price prediction with Convolutional neural network and Stacked Gated Recurrent Unit. *Data, 7*(11), 149. https://doi.org/10.3390/data7110149

Kilimci, Z. H. (2020). Sentiment analysis based direction prediction in bitcoin using deep learning algorithms and word embedding models. *International Journal of Intelligent Systems and Applications in Engineering, 8*(2), 60–65. https://doi.org/10.18201/ijisae.2020261585

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Sage.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (4th ed.). SAGE Publications.

Lahmiri, S., & Bekiros, S. (2020). The impact of COVID-19 pandemic upon stability and sequential irregularity of equity and cryptocurrency markets. *Chaos, Solitons & Fractals, 138*, 109936. https://doi.org/10.1016/j.chaos.2020.109936

Lamon, C., Nielsen, E., & Redondo, E. (2017). Cryptocurrency price prediction using news and social media sentiment. *SMU Data Science Review, 1*, 1–22.

Li, X., Ding, L., Du, Y., Fan, Y., & Shen, F. (2022). Position-enhanced multi-head self-attention based bidirectional gated recurrent unit for aspect-level sentiment classification. *Frontiers in Psychology, 12*, 799926. https://doi.org/10.3389/fpsyg.2021.799926

Lisivick, M. (2017). *NewsAPI* [Computer software]. https://github.com/mattlisiv/newsapi-python

Liu, X., Zhou, G., Kong, M., Yin, Z., Li, X., Yin, L., & Zheng, W. (2023). Developing multi-labelled corpus of twitter short texts: A semi-automatic method. *Systems, 11*(8), 390. https://doi.org/10.3390/systems11080390

Loginova, E., Tsang, W. K., van Heijningen, G., Kerkhove, L.-P., & Benoit, D. F. (2021). Forecasting directional bitcoin price returns using aspect-based sentiment analysis on online text data. *Machine Learning*. https://doi.org/10.1007/s10994-021-06095-3

Loughran, T., & Mcdonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance, 66*(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Loughran, T., & Mcdonald, B. (2014). Measuring Readability in financial disclosures. *The Journal of Finance, 69*(4), 1643–1671. https://doi.org/10.1111/jofi.12162

Luo, J. (2020). Bitcoin price prediction in the time of COVID-19. *2020 Management Science Informatization and Economic Innovation Development Conference (MSIEID)* (pp. 243–247).https://doi.org/10.1109/MSIEID52046.2020.00050

Mai, F., Shan, Z., Bai, Q., Wang, X., & Chiang, R. H. L. (2018). How does social media impact Bitcoin value? A test of the Silent Majority Hypothesis. *Journal of Management Information Systems, 35*(1), 19–52. https://doi.org/10.1080/07421222.2018.1440774

Manaf, S. A., & Nordin, M. J. (2009). Review on statistical approaches for automatic image annotation. *2009 International Conference on Electrical Engineering and Informatics* (pp. 56–61).https://doi.org/10.1109/ICEEI.2009.5254815

Maqsood, U., Khuhawar, F. Y., Talpur, S., Jaskani, F. H., & Memon, A. A. (2022). Twitter Mining based Forecasting of cryptocurrency using sentimental analysis of Tweets. *2022 Global Conference on Wireless and Optical Technologies (GCWOT)* (pp. 1–6).https://doi.org/10.1109/GCWOT53057.2022.9772923

Mohammad, S. (2016). A Practical Guide to Sentiment Annotation: Challenges and Solutions. *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 174–179). https://doi.org/10.18653/v1/W16-0429

Mohanty, P., Patel, D., Patel, P., & Roy, S. (2018). Predicting fluctuations in cryptocurrencies' price using users' comments and real-time prices. *7th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO).*

Mohapatra, S., Ahmed, N., & Alencar, P. (2019). KryptoOracle: A real-time cryptocurrency price prediction platform using Twitter sentiments. *2019 IEEE International Conference on Big Data (Big Data)* (pp. 5544–5551).https://doi.org/10.1109/BigData47090.2019.9006554

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. *Decentralized Business Review* 21260.

Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining, 11*(1), 81. https://doi.org/10.1007/s13278-021-00776-6

Nasekin, S., & Chen, C.Y.-H. (2020). Deep learning-based cryptocurrency sentiment construction. *Digital Finance, 2*(1–2), 39–67. https://doi.org/10.1007/s42521-020-00018-y

Oikonomopoulos, S., Tzafilkou, K., Karapiperis, D., & Verykios, V. (2022). Cryptocurrency Price Prediction using Social Media Sentiment Analysis. *2022 13th International Conference on Information, Intelligence, Systems & Applications (IISA)* (pp. 1–8). https://doi.org/10.1109/IISA56318.2022.9904351

Ortu, M., Uras, N., Conversano, C., Bartolucci, S., & Destefanis, G. (2022). On technical trading and social media indicators for cryptocurrency price classification through deep learning. *Expert Systems with Applications, 198*, 116804. https://doi.org/10.1016/j.eswa.2022.116804

Pant, D. R., Neupane, P., Poudel, A., Pokhrel, A. K., & Lama, B. K. (2018). Recurrent neural network based Bitcoin price prediction by Twitter sentiment analysis. *2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS)* (pp. 128–132). https://doi.org/10.1109/CCCS.2018.8586824

Parekh, R., Patel, N. P., Thakkar, N., Gupta, R., Tanwar, S., Sharma, G., Davidson, I. E., & Sharma, R. (2022). DL-GuesS: Deep learning and sentiment analysis-based cryptocurrency price prediction. *IEEE Access, 10*, 35398–35409. https://doi.org/10.1109/ACCESS.2022.3163305

Passalis, N., Avramelou, L., Seficha, S., Tsantekidis, A., Doropoulos, S., Makris, G., & Tefas, A. (2022). Multisource financial sentiment analysis for detecting Bitcoin price change indications using deep learning. *Neural Computing and Applications, 34*(22), 19441–19452. https://doi.org/10.1007/s00521-022-07509-6

Pathak, S., & Kakkar, A. (2020). Cryptocurrency price prediction based on historical data and social media sentiment analysis. *Proceedings of 7th Innovations in Computer Science and Engineering (ICICSE), 97*, 177–192.

Pillai, S., Biyani, D., Motghare, R., & Karia, D. (2021). Price prediction and notification system for cryptocurrency share market trading. *2021 International Conference on Communication Information and Computing Technology (ICCICT)* (pp. 1–7). https://doi.org/10.1109/ICCICT50803.2021.9510122

Pintelas, E., Livieris, I. E., Stavroyiannis, S., Kotsilieris, T., & Pintelas, P. (2020). Investigating the problem of cryptocurrency price prediction: A deep learning approach. In I. Maglogiannis, L. Iliadis, & E. Pimenidis (Eds.), *Artificial Intelligence Applications and Innovations* (Vol. 584, pp. 99–110). Springer International Publishing.

Prajapati, P. (2020). Predictive analysis of Bitcoin price considering social sentiments. arXiv:2001.10343. http://arxiv.org/abs/2001.10343

Qiao, Y., Xiong, C., Liu, Z., & Liu, Z. (2019). *Understanding the Behaviors of BERT in Ranking* (arXiv:1904.07531). http://arxiv.org/abs/1904.07531

Raju, S. M., & Tarif, A. M. (2020). *Real-time prediction of Bitcoin price using machine learning techniques and public sentiment analysis* 14.

Reis, J., Benevenuto, F., Olmo, P., Prates, R., Kwak, H., & An, J. (2015). *Breaking the News: First Impressions Matter on Online News*.

Riccosan, & Saputra, K. E. (2023). Multilabel multiclass sentiment and emotion dataset from indonesian mobile application review. *Data in Brief, 50*, 109576. https://doi.org/10.1016/j.dib.2023.109576

Ritchie, M. J., Drummond, K. L., Smith, B. N., Sullivan, J. L., & Landes, S. J. (2022). Development of a qualitative data analysis codebook informed by the i-PARIHS framework. *Implementation Science Communications, 3*(1), 98. https://doi.org/10.1186/s43058-022-00344-9

Rognone, L., Hyde, S., & Zhang, S. S. (2020). News sentiment in the cryptocurrency market: An empirical comparison with Forex. *International Review of Financial Analysis, 69*, 101462. https://doi.org/10.1016/j.irfa.2020.101462

Salač, A. (2019). *Forecasting of the cryptocurrency market through social media sentiment analysis* [[Student Theses], University of Twente]. https://essay.utwente.nl/78607/

Sarica, S., & Luo, J. (2021). Stopwords in technical language processing. *PLoS ONE, 16*(8), e0254937. https://doi.org/10.1371/journal.pone.0254937

Sattarov, O., Jeon, H. S., Oh, R., & Lee, J. D. (2020). Forecasting Bitcoin Price Fluctuation by Twitter Sentiment Analysis. *2020 International Conference on Information Science and Communications Technologies (ICISCT)* (pp. 1–4).https://doi.org/10.1109/ICISCT50599.2020.9351527

Schulte, M., & Eggert, M. (2021). *Predicting hourly bitcoin prices based on long short-term memory neural networks*.

Serafini, G., Yi, P., Zhang, Q., Brambilla, M., Wang, J., Hu, Y., & Li, B. (2020). Sentiment-driven price prediction of the Bitcoin based on statistical and deep learning approaches. *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). https://doi.org/10.1109/IJCNN48605.2020.9206704

Seroyizhko, P., Zhexenova, Z., Shafiq, M. Z., Merizzi, F., Galassi, A., & Ruggeri, F. (2022). A sentiment and emotion annotated dataset for bitcoin price forecasting based on reddit posts. *Proceedings of the 4th Workshop on Financial Technology and Natural Language Processing*.

Shah, N., & Rohilla, S. (2022). *Emot* (3.1) [Computer software]. https://github.com/NeelShah18/emot

Shahzad, M. K., Bukhari, L., Khan, T. M., Islam, S. M. R., Hossain, M., & Kwak, K.-S. (2021). BPTE: Bitcoin Price Prediction and Trend Examination using Twitter Sentiment Analysis. *2021 International Conference on Information and Communication Technology Convergence (ICTC)* (pp. 119–122).https://doi.org/10.1109/ICTC52510.2021.9620216

SNScrape. (2018). [Computer software]. https://github.com/JustAnotherArchivist/snscrape

Steinert, L., & Herff, C. (2018). Predicting altcoin returns using social media. *PLoS ONE, 13*(12), e0208119. https://doi.org/10.1371/journal.pone.0208119

Stenqvist, E., & Lönnö, J. (2017). *Predicting Bitcoin price fluctuation with Twitter sentiment analysis* [Degree Project]. KTH Royal Institute of Technology School of Computer Science and Communication.

Sukumaran, S., Bee, T. S., & Wasiuzzaman, S. (2022). Cryptocurrency as an investment: The Malaysian context. *Risks, 10*(4), 86. https://doi.org/10.3390/risks10040086

Symeonidis, S., Kordonis, J., Effrosynidis, D., & Arampatzis, A. (2017). DUTH at SemEval-2017 Task 5: Sentiment predictability in financial microblogging and news articles. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 861–865). https://doi.org/10.18653/v1/S17-2147

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics, 37*(2), 267–307. https://doi.org/10.1162/COLI_a_00049

Valencia, F., Gómez-Espinosa, A., & Valdés-Aguirre, B. (2019). Price movement prediction of cryptocurrencies using sentiment analysis and machine learning. *Entropy, 21*(6), 589. https://doi.org/10.3390/e21060589

Van Atteveldt, W., Van Der Velden, M. A. C. G., & Boukes, M. (2021). The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures, 15*(2), 121–140. https://doi.org/10.1080/19312458.2020.1869198

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is All you Need*.

Vo, A.-D., Nguyen, Q.-P., & Ock, C.-Y. (2019). Sentiment analysis of news for effective cryptocurrency price prediction. *International Journal of Knowledge Engineering, 5*(2), 47–52. https://doi.org/10.18178/ijke.2019.5.2.116

Wan, X., Yang, J., Marinov, S., Calliess, J.-P., Zohren, S., & Dong, X. (2021). Sentiment correlation in financial news networks and associated market movements. *Scientific Reports, 11*(1), 3062. https://doi.org/10.1038/s41598-021-82338-6

Wołk, K. (2019). Advanced social media sentiment analysis for short-term cryptocurrency price prediction. *Expert Systems, 37*(2), 1–16. https://doi.org/10.1111/exsy.12493

Wooley, S., Edmonds, A., Bagavathi, A., & Krishnan, S. (2019). Extracting cryptocurrency price movements from the Reddit network sentiment. *2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 500–505). https://doi.org/10.1109/ICMLA.2019.00093

Yang, Y., Zha, K., Chen, Y.-C., Wang, H., & Katabi, D. (2021). Delving into deep imbalanced regression. *International Conference on Machine Learning*.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2020). XLNet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems (NeurIPS 2019)* (pp. 1–11).

Yao, W., Xu, K., & Li, Q. (2019). Exploring the influence of news articles on Bitcoin price with machine learning. *2019 IEEE Symposium on Computers and Communications (ISCC)* (pp. 1–6).https://doi.org/10.1109/ISCC47284.2019.8969596

Yusof, M. F., Ab. Rasid, L., & Masri, R. (2021). Implementation Of Zakat Payment Platform For Cryptocurrencies. *AZKA International Journal of Zakat & Social Finance*. https://doi.org/10.51377/azjaf.vol2no1.41

Zamani, N. A. M., & Kamaruddin, N. (2023). Crypto-sentiment detection in Malay text using language models with an attention mechanism. *Journal of Information System Engineering and Business Intelligence, 9*(2), 147–160.

Zamani, N. A. M., Liew, J. S. Y., & Yusof, A. M. (2022a). XLNET-GRU sentiment regression model for cryptocurrency news in English and Malay. *Proceedings of the 4th Financial Narrative Processing Workshop @ LREC 2022* (pp. 36–42).

Zamani, N. A. M., Yan, J. L. S., & Yusof, A. M. (2022b). Cryptocurrency price prediction using Bi-GRU model with English and Malay news sentiment features. *2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS)* (pp. 136–141). https://doi.org/10.1109/AiDAS56890.2022.9918725

Zhang, X., Wu, Z., Liu, K., Zhao, Z., Wang, J., & Wu, C. (2023). Text sentiment classification based on BERT embedding and sliced multi-head self-attention bi-GRU. *Sensors*. https://doi.org/10.3390/s23031481

Zolkepli, H. (2022). *Malaya Documentation* [Computer software]. https://malaya.readthedocs.io/en/stable/Dataset.html#

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Nur Azmina Mohamad Zamani[1,2] · **Norhaslinda Kamaruddin**[3] · **Ahmad Muhyiddin B. Yusof**[4]

✉ Nur Azmina Mohamad Zamani
namz.ina@gmail.com; azmina@uitm.edu.my

Norhaslinda Kamaruddin
norhaslinda@tmsk.uitm.edu.my

Ahmad Muhyiddin B. Yusof
ahmadmuhyiddin4@uitm.edu.my

[1] College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

[2] College of Computing, Informatics and Mathematics, Universiti Teknologi MARA Perak Branch, Tapah Campus, 35400 Tapah Road, Perak, Malaysia

[3] Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Universiti Teknologi MARA (UiTM), Kompleks Al-Khawarizmi, 40450 Shah Alam, Selangor, Malaysia

[4] Academic of Language Studies, Pusat Asasi, Universiti Teknologi MARA Selangor Campus, 43800 Dengkil, Selangor, Malaysia