The Impact of Different Social Media Sentiments on Bitcoin Price Forecasting

ABSTRACT

Cryptocurrencies, led by Bitcoin, together form an emerging market for trading and investment, and because of their high volatility, their price forecasts are highly instructive and difficult. Therefore, this paper proposes a approach to collect social media user sentiments and utilize them as one of the factors to verify their impact on Bitcoin price forecasts on different models (LSTM, GBDT); it also provides a brief comparison of the extent to which different styles of social medias affect prices.

CCS CONCEPTS

• **Information systems** → **Web mining.**;

KEYWORDS

neural networks, sentiment analysis, social media, time series, forecasting

1 INTRODUCTION

Bitcoin is an electronic currency generated by the open-source P2P software, and is currently the most widely used cryptocurrency in the world. More and more investors are choosing Bitcoin as their primary investment product, and as a direct result, Bitcoin's market capitalization has skyrocketed. As a virtual commodity generated by a specific computer program, the price of Bitcoin is not only influenced by the law of supply and demand, but also by the increasing cost of production, risk aversion and speculative demand. The Bitcoin market is a highly speculative and volatile market, and it is inefficient compared to other investment markets. Therefore, the study of Bitcoin price prediction and analysis to better determine the future trend of Bitcoin is of great personal and social significance.

Traditional financial models based entirely on their counterpart markets have a well established development for forecasting stock prices, however, these models often yield less than optimal results when forecasting Bitcoin prices. While traditional financial markets price assets such as stocks, currencies and gold based on certain fundamental assets, the complexity of the price formation mechanism of digital currencies, the extremely non-linear, non-stationary and complex nature of their prices, as well as the decentralized and unregulated nature, increase the difficulty of analyzing and forecasting their prices.

With the various nascent social media outlets popping up today, they have also become a marketing tool for cryptocurrencies, led by Bitcoin, as public discussions about cryptocurrencies are taking place on social media channels. This chatter can be monitored and used to predict cryptocurrency prices based on public sentiment. Meanwhile, Bitcoin's price is influenced by financial institutions and celebrities in the same way as stocks, if not more so. For example, on January 19, 2021, Elon Mask, co-founder of Paypal and CEO of Tesla, added Bitcoin to his Twitter page and tweeted "In retrospect, it was inevitable", leading to Bitcoin's price climbed $5,000 to $37,299 within an hour.

So, is it possible to analyze the sentiment about Bitcoin on social media and forecast the price of Bitcoin by means of natural language processing? This depends on whether there is a correlation between social media data and Bitcoin price fluctuations. Moreover, users interact differently on different types of social media, so are the sentiments we capture the same or different, and to what extent do they each reflect the Bitcoin prices? We crawled user contribution data from two social media outlets and examined the questions in conjunction with historical Bitcoin prices to verify them in practice.

2 RELATED WORK

Financiers have stated that decisions in the financial system are influenced by emotional ethics, not just capital values. Dolan and Edlin[4] supports this view and argues that decisions are influenced by emotions. Therefore, tools such as sentiment analysis help to show that commodity prices are influenced by values such as emotions and economic fundamentals. In the following we will divided into two modules based on our paper.

Sentiment analysis. Dealing with the emotion of a text is already a more mature field. lexicon based models are considered the most classical language models, semantic orientation represents polarity and strength of a positive/negative word and is a measure of subjectivity and opinion in text. Moreover, Jurek et al.[1] proposed a lexicon-based model for sentiment analysis of social networks, and they have been used to estimate sentiment strength rather than just positive/negative labels. In the direction of machine learning and deep learning, there are also many higher performance models emerging, such as Hidden Markov Model, Recursive Neural Network (RNN). The emergence of pre-trained language models has opened a new chapter. large language models have demonstrated better performance in real-world sentiment analysis. Singh et al.[8] obtained 94% accuracy on tweets data with pre-trained Bidirectional Encoder Representations from Transformers (BERT) models. In this paper we adapted both lexicon-based models updated with financial lexicon and BERT models to obtain sentiment scores

Price Forecasting. It can be considered as a sub-task of time series forecasting. Box and Jenkins, in the late 70s, made an important work in studying applications composed of mathematical linear models. This method became a classical algorithm later on, however, it does not perform well in non-linear behavior predictions. Meanwhile, some studies based on Artificial Neural Network have progressed in recent years, Oancea and Ciucu[2] built a neural network with RNN architecture to predict exchange rate movements and compared it with classical financial models (BX-Jenkin, ARIMA) to illustrate the effectiveness of neural network architecture for price prediction. Lara-Ben´ıtez et al.[5] compare the neural network models that have emerged in recent years and conclude that the Long Short-Term Memory (LSTM) model has the best performance among the various neural network architectures. Also in

Table 1: Subreddit Statistics

Comments	Members	Submissions
CryptoCurrency	4.4m	712k
1.6b		
CryptoMarket	1.4k	97k
228m		
Bitcoin	3.8m	566k
1.1m		
btc	780k	66k
183m		

Table 2: Number of Submissions

All Data	Valid Data	BTC-related Data
1,097,120	764,267	247,176

the field of machine learning, chen et al.[3] obtained high accuracy in predicting prices by tree models and gradient boosting algorithm, which proved its feasibility. In this paper we take LSTM and GBDT models for verification.

3 DATASET PREPROCESSING

Our data are mainly derived from two social medias, Reddit, which represents the forum type of social media, and Telegram, which represents the group chat type of social media. They are discussed separately below.

Reddit. Users interact in two main ways. The first way is by posting submissions, where they publish a new topic and get other users' thoughts on it; the second way is by posting comments, where they share their direct thoughts on existing threads, or reply to other people's comments on that. To obtain user contributions related only to the cryptocurrency market, we focused on *CryptoCurrency, Bitcoin, btc, and CryptoMarket subreddits*, and we obtained all user contributions on these subreddits for the two years from 2020 to 2021, totaling more than one million entries, see Table 1.

First, we need to crawl the textual and numerical data of each submission on Reddit from the web. We use PMAW to fetch the data from *PushShift* asynchronously, and PRAW to assist by updating some of the outdated data. In the obtained submissions, we keep the timestamp, the ratio of up-votes, the number of up-votes, and the number of comments in addition to the text information. Specially, we remove all comments from our dataset because the heterogeneity between the comments and the submissions is such that they interfere with the computation of the sentiment score for the whole submission[7], and we take another approach to compute the degree of affirmation of this submission by other users in the later section.

Telegram. When a user joins a group, he or she is free to post SMSes with a certain word limit, either in response to a previous user or to start a new topic, and in this paper we apply message to refer to user contributions in group chat social media. We selected several English-speaking groups that are widely acknowledged in the cryptocurrency market and have a high number of users and activeness; we obtained a total of 857,267 messages from February 2021 to the end of 2022, using official methods. They were parsed and converted into the same data format as Reddit.

Preprocessing. For the data obtained from these two sources, we mark them as forum data and group chat data respectively. For the forum data, in order to focus on the valid submissions from the actual users, we first removed the illegal data, filtered out the deleted and removed submissions by authors or forum admins, and removed the submissions identified as bots by the system or with "bot" in the author's name. For all the submissions, we analyzed them and found that only about 24% of them were about Bitcoin, see Table 2. In order to ensure that the topics of the forum and group chat messages correspond to each other, we finally used a broader set of all cryptocurrency market data, which holds 76% of the original data. For the group chat data, we removed invalid information (e.g., links, images, contact information, spam words, etc.)

4 POST SENTIMENT

In this section we will perform sentiment scoring on the obtained single post. This scoring is to be applied to both forum data and group chat data. At the beginning, we rate the text of a post in terms of their positive and negative level, and after that, we score the post based on its degree of popularity and other users' approval of the post to get a combined sentiment score of the post.

4.1 Text Polarities

We employ the combined polarities of two models to obtain the sentiment carried by a piece of text. The first linguistic model is VADER , which built on a lexicon-based and rule-based model, and it has a lexicon attuned to sentiment expressed by social media . In particular, we updated the lexicon of VADER with Loughran McDonald's word list, which internally assigns new weights to more than four thousand finance-related words (354 positive words and 2355 negative words chosen) to make it more suitable for text analysis in the cryptocurrency market. The second model adapts a deep learning algorithm, based on the BERT architecture, and we reuse an open-source model weights from *HuggingFace*. This model was trained on the Stanford Treebank dataset and is pre-trained and fine-tuned on three architectures, BERT, AlBERT and DistilBERT. for each of these two models, we collected sentiment polarities and plotted the distribution on the same data sampling.

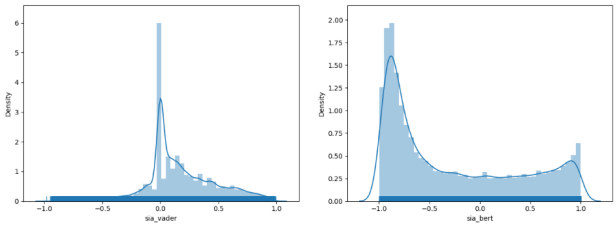

Figure 1: Distribution of polarities by VADER and BERT, ranging from -1 to +1.

Here we map the polarity values of both models to the range between -1 and +1, with higher values representing more positive sentiment of the text. From Fig 1, it can be observed that VADER generates conservative polarity scores, with most scores clustered around 0, while BERT tends to offer very strong conclusions, and its scores are also more evenly distributed within the interval. While VADER is based on a more traditional language model, it can capture financial information well after lexicon updates; whereas BERT, although not trained specifically for financial texts, its weights obtained based on large-scale text training have been shown to achieve good results on a variety of generalized text applications . By weighting and averaging these two models, which are based on different principles and have different scoring modes, we can obtain a more accurate text sentiment polarity score that incorporates the advantages of both models.

4.2 Confidence Level

After getting the sentiment score of the text in a post, we also need to get the confidence level of a post through other attributes, and then calculate the sentimental tendency score of it. Here we set the following formula

$$sentiment\_score = p\_scale \times c\_scale \times text\_polarity \quad (1)$$

where popularity scale indicates the popularity of the post. Its value ranges from 0.5 to 1, and higher value represents more popularity, the scale is derived from the following formula

$$p\_scale = sigmoid(n_s + w \times n_c) \quad (2)$$

where w is the transformed coefficient between number of votes and number of comments. In the forum data, there exists a quantitative relationship between these two, and we identified the value of w by regression analysis upon the complete forum data. In the group chat data, w defaults to 0, while number of votes is 1, indicating that only the person who sent the message agrees with itself; number of comments is 0, implying that no one has made any comment on this message. We normalize the summation with the sigmoid function to ensure that the popularity scale remains at least 0.5 even when there are few comments and votes, and does not exceed 1 also when a post goes viral.

Condifence scale is derived from the following formula, and takes values from -1 to +1, with -1 indicating strong disagreement and +1 indicating fully agreement. The default value in the group chat data is 1, which means that only the person who sent the message gave itself an up-vote and no one gave a down-vote.

$$c\_scale = 2 \times upvote\_ratio - 1 \quad (3)$$

By multiplying (2) and (3) with the original text polarity, we can ultimately obtain the overall score of a post on social media. Note that here the chat message can be brought into the formula as a corner case of the forum message.

5 PRICE FORECASTING

In this section we will first analyze the correlation between sentiment and the current day's Bitcoin price and its validity for predicting the price of Bitcoin. After that, we will include it as one of the indicators in two models for forecasting the prices of Bitcoin.

5.1 Correlation Analysis

In order to process the post sentiment score into a format compatible with other financial data, we first aggregate each post according to its date to obtain a composite daily sentiment score. To determine the relationship between daily sentiment scores and prices, we obtained the closing prices of Bitcoin and other relevant technical indicators from *TradingView* for the past year. We calculated the Pearson correlation coefficients for these indicators and plotted the heat map.

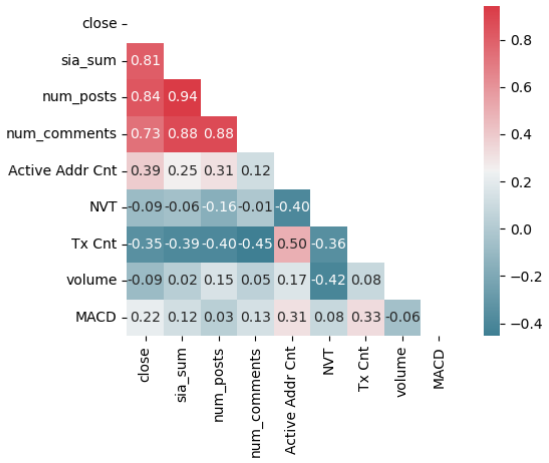

Figure 2: Pearson correlation coefficient of features and target (close) at the first column.

From the diagram it can conclude that the daily sentiment score and the daily closing price show a very high correlation, even higher than many of the technical indicators. However, when applying the next day's coin price, the correlation becomes very poor. This illustrates two points, firstly the daily sentiment score we calculated before does reflect the price of the day; but if employed for the next day's price forecasting, a simple model may not perform well and we need a more advanced model to support the forecasting task.

5.2 Feature Selection & Preprocessing

We selected several representative features in fields of finance and blockchain to expand the information content of the input features, also considering the the correlation of them and the target in the previous section. The feature set was decided to be,

- `Daily Sentiment Score`: The aggregated score from the sentiment calculated in the last section.
- `Number of Posts`: The number of submissions/messages submitted per day.
- `Volume`: The number of traded shares.

- `Exponential Moving Average (EMA)`: Weighted average which placing a greater weight and significance on the most recent daily prices.
- `Moving Average Convergence Divergence (MACD)`: The difference between an instrument's 26-day and 12-day EMA.
- `Relative Strength Index (RSI)`: A momentum oscillator which measuring the speed and change of price movements.
- `Network Value to Transaction (NVT)`: Describing the relationship between transfer volume and market capitalization.
- `Active Address Count`: The count of addresses with transactions.
- `Block Size`: Implying the the amount of transactions it stores.

To prevent the daily price from fluctuating too much, we took the natural logarithm of the price. The last 30 days of data are collected as the test set and all the previous data are considered as the training set. We also did feature scaling on both sets for all the input values.

5.3 LSTM Model

RNN is a classical network architecture for time series analysis, however, it is hampered by short-term memory. If a sequence is long enough, it will be tough for them to transfer information from an earlier time step to a later time step. Therefore, if a sequence of data is processed for prediction, RNN may miss important information from the very beginning. During backpropagation, RNN faces the problem of gradient disappearance and stops learning if the gradient value becomes very small. A solution to the short term memory problem is the LSTM. It has been shown to be particularly effective for learning sequences containing long-term patterns of unknown length because of its ability to maintain long-term memory and also has shown good performance on many time series prediction tasks . The Bitcoin market, while it is difficult to capture seasonal price variability, has a unique chart pattern when the market as a whole is bullish or bearish , which makes the choice to utilize LSTM justifiable.

We built a neural network with double LSTM layers, each layer having 50 cells, followed by dropout and sense layers. for every of the target, we took as input all the data containing the first 30 days of the target, that is, a matrix with a shape of 30 by 9. We employed cross validation because the input data of this network is small. We applied grid search to find the hyperparameters and finally determined the best model when the batch size is 8, number of epochs is 20 and optimizer is ADAM. Figure 3 shows the test scores of the model trained in each step of grid search.

5.4 GBDT Model

The basic concept of Gradient Boosting is to sequentially generate multiple weak learners, each with the goal of fitting the negative gradient of the loss function of the previously accumulated model, such that the loss of the accumulated model after adding the weak learner decreases in the direction of the negative gradient. It is an algorithm with strong generalization ability. The idea offers a natural advantage to find a variety of distinguishing features.

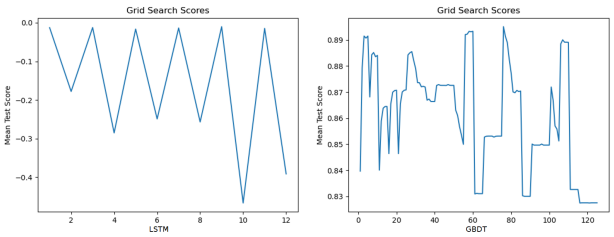

Figure 3: Test score of LSTM model and GBDT in grid search.

The combination of Decision Tree and Gradient Boosting led to the Gradient Boosting Decision Tree (GBDT), which inherits many advantages of decision trees while improving their disadvantages. As the trees adopted in GBDT are all low-complexity trees, the variance becomes tiny, and the overfitting problem can be effectively handled by integrating multiple decision trees through the gradient boosting method.

We built the algorithm with XGBoost . XGBoost provides a scalable end-to-end tree boosting system. It not only introduces regular terms into the loss function, but also defines a criterion for tree node splitting - using 1st and 2nd order gradients to determine the optimal splitting point - which has been proven effective in engineering practices. Again, we selected the feature data of the first 30 days and flattened them to get a vector of length 270 as inputs. Cross validation and grid search were applied to find the hyperparameters, and finally located the best model when the learning rate is 0.4, max depth is 1, and number of estimators is 20. Figure 3 illustrates the model test scores of each step of grid search.

5.5 Result

For the two models trained above, we performed forecasts on each of the two types of social media data. The forecasting results and the scores are shown in Figure 4. To determine the validity of these two models, we employed the model that repeated the previous day's coin price as the baseline, whose unsquared MSE in the test set was 0.042073, while all four of our models outperformed the baseline.

It is observed that GBDT outperforms the results of LSTM on all datasets tested, but the forecasting curve of the LSTM model is relatively smooth and presents better continuity. We also found that models trained from group chat data were overall preferred to those trained using forum data. This also distinguishes, to some extent, the varying effects of different social media on price forecasts.

6 DISCUSSION

From the above, it follows that firstly, user sentiment on social media does have an impact on the current day's coin price; and it is also feasible to predict the future price by utilizing historical user sentiment data. In this experiment, the models were also trained separately in different forms of social media datasets, and the group chat data outperformed the forum data in both different model architectures for coin price forecasting. The data from the group chat, although relatively more casual, nevertheless presents a high value for price forecasts

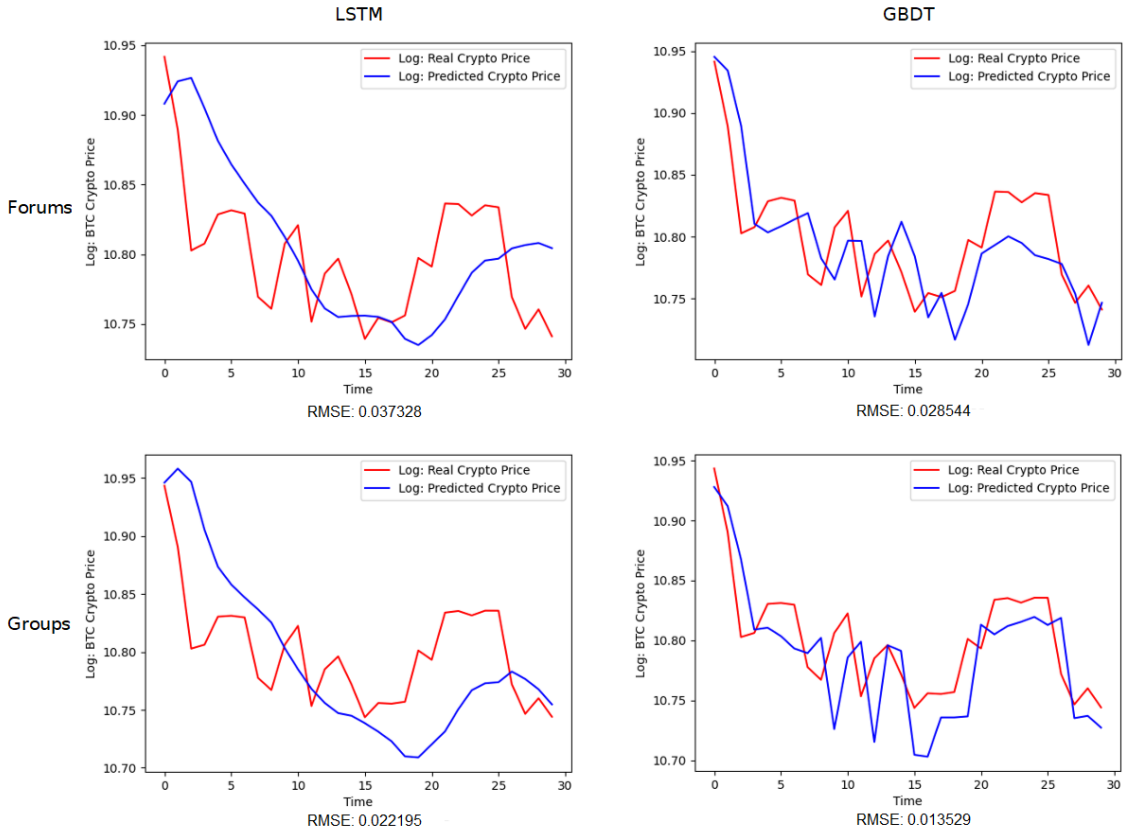

Figure 4: Prediction curves of LSTM/GBDT with forum/group data.

Figure 5: Word cloud generated from forum and group chat data.

We analyzed the frequency of occurrence of specific words in both datasets through word clouds. As we can see, "crypto" and "Bitcoin" are frequently mentioned in the forum data, while the most frequently mentioned term in the chat room data is "project ". This again highlights that people tend to communicate in different styles across social media. Although both are cryptocurrency market related forums or groups, there are differences in the subjects that users discuss on the two social media outlets. This is probably one of the reasons why the group chat media performs better in price forecasting.

The characteristics of social media are complex, and the different characteristics are distinguished by various dimensions such as directionality, synchronicity, etc .[6] In terms of behavior patterns, the way between forum users is dual, users will discuss under a certain forum submission and vote for this in the meantime. Whereas in group chats, users send messages in higher synchronicity, and they do not vote for other people's messages.In terms of topics, forum discussions are all under one post, and users will leave topic-related comments under a specific submission, while a submission that does not interest the user will be replaced by another submission; whereas group chat discussions do not have a specific topic, and users can reply to any of the previous messages or start a new discussion. The style of writing is also different in this setting, as forum users express themselves in a more formal tone, with relatively few omitted subjects and objects, and the length of each message is longer, while group chat users express themselves in a more casual style, with many omissions and colloquialisms, and each message is relatively short . These characteristics work together to render the media as a whole with a different sentiment tone, thus showing different valuation on the task of Bitcoin price forecasting.

7 CONCLUSION & FUTURE WORK

This paper explores the relationship between the overall sentiment of different types of social media and Bitcoin price, proposes a general algorithm for computing post sentiment, and finally validates

the Bitcoin price on different models. The GBDT model in this paper slightly outperforms the LSTM model in terms of scoring, while the LSTM in this paper may also further improve its performance after more elaborate tuning. Meanwhile, the overall prediction accuracy may also be improved by selecting other technical indicators or adopting other time series prediction models. In terms of social media, the data from group chat media shows higher correlation comparing to forum media; however, collecting more media of the same type for a comprehensive comparison would strengthen the persuasiveness of this conclusion.

REFERENCES

[1] Maurice D. Mulvenna Anna Jurek and Yaxin Bi. 2015. Improved lexicon-based sentiment analysis for social media analytics. *Security Informatics* 4, 9, Article 4 (Nov 2015). https://doi.org/10.1186/s13388-015-0024-x

[2] ŞTefan Cristian Ciucu Bogdan Oancea. 2017. Time series forecasting using neural networks. *Proceedings of the CKS 2013 International Conference* (Jan. 2017). https://arxiv.org/abs/1401.1333

[3] Zheshi Chen, Chunhong Li, and Wenjun Sun. 2020. Bitcoin price prediction using machine learning: An approach to sample dimension engineering. *J. Comput. Appl. Math.* 365 (2020), 112395. https://doi.org/10.1016/j.cam.2019.112395

[4] Paul Dolan and Richard Edlin. 2002. Is it really possible to build a bridge between cost-benefit analysis and cost-effectiveness analysis? *Journal of Health Economics* 21, 5 (2002), 827–843. https://doi.org/10.1016/S0167-6296(02)00011-5

[5] José C. Riquelme Pedro Lara-Benítez, Manuel Carranza-García. 2021. An Experimental Review on Deep Learning Architectures for Time Series Forecasting. *International Journal of Neural Systems* 31, 3 (April 2021).

[6] Mason-R. Romiszowski, A. J. 1996. Computer-mediated communication. In D. Jonassen (Ed.). *Handbook of research for educational communications and technology* (1996), 438–456.

[7] Daniel Röchert, German Neubaum, Björn Ross, Florian Brachten, and Stefan Stieglitz. 2020. Opinion-based Homogeneity on YouTube: Combining Sentiment and Social Network Analysis. *Computational Communication Research* 2 (02 2020), 81–108. https://doi.org/10.5117/CCR2020.1.004.ROCH

[8] Jakhar A.K. Singh, M. and S. Pandey. 2021. Sentiment analysis on the impact of coronavirus in social life using the BERT model. *Social Network Analysis and Mining* 11, 33 (March 2021), 36–44. https://doi.org//10.1007/s13278-021-00737-z