Predicting Trends in Bitcoin Prices Using Twitter Sentiment Analysis

December 2021

AJ Valenty
Data Science
Rice University
Houston Texas USA
amv10@rice.edu

Ziyana Samanani
Data Science
Rice University
Houston Texas USA
zs18@rice.edu

ABSTRACT

This project aims to analyze the relationship between tweet sentiment and Bitcoin price trends to see if public sentiment can function as a predictor for the Bitcoin price for time intervals within a day. Social media has the power to influence many current events, and we wonder if we can use the general public's sentiment to predict the Bitcoin price. Twitter sentiment and Bitcoin price data, labeled based on the direction of price change and closing price, were analyzed as both a prediction and classification, testing both Long Short Term Memory and Logistic Regression models. Overall results varied widely depending on time interval and amount of data. The final models with the best performance were the 1-hour LSTM, having an R^2 of 0.63 and MSE of 0.06, and the 6-hour logistic regression, having an accuracy of 0.557. Due to our dataset and past work, we determine that it is still quite hard to accurately determine the Bitcoin price or difference given solely twitter sentiment, and that sentiment is more of a lagging indicator instead of a direct indicator.

1 Introduction

Cryptocurrency is a decentralized transaction system that uses cryptography techniques to cut out middlemans like banks and governments to make transactions, helping problems like inflation and corruption. Cryptocurrencies have been gaining popularity, as Tesla purchased $1.5 billion dollars worth of Bitcoin in February of this year and was accepting Bitcoin as a payment method [1] . Since the start of the global pandemic caused by COVID-19, the buying and selling of cryptocurrency has become increasingly popular for growing wealth [2], given its 60% increase in price in the last year. There has also been growth in the young retail investors - young people with available funds that they want to invest and earn a profit. According to recent research, 68% of American millionaires are invested in cryptocurrency and 54% of them use social media as their main source of cryptocurrency information [3]. As a result, social media is an incredibly powerful tool that high-impact crypto investors use as a factor for gauging whether to buy or sell cryptocurrency. It would be helpful for individuals to have information on how much social media can persuade the price, assisting them in making smart decisions with their money. Additionally, the ability to make profits using unorthodox trading strategies in this new form of investment is incredibly lucrative.

Online forums including Twitter, Discord and Reddit feature many discussions about crypto-related trading moves and technical analysis. These communities can cause massive changes in the U.S. market, as with the GameStop short squeeze in early 2021 resulting in a 3000% jump in its stock price, but what about in the unregulated market of cryptocurrency? Many celebrities and influencers are being paid to market specific altcoins, without having any knowledge about what they are supporting, being known as pump and dump* schemes. Consequently, it is important to determine if there is a relationship between the public perception of a coin and its price, and with that, if a shift in perception causes a change in price.

We are curious whether we can use social media sentiment as the sole predictor of the price. While prior work has found a strong correlation between *interday* social media sentiment and crypto price, the goal of this project is to see if the same is true for *intraday* sentiment. More specifically, we want to understand the relationship between the sentiment of Bitcoin-related tweets from different time intervals in a single day and the corresponding fluctuation in this coin's price over the same time period. With this knowledge, we can assess if analyzing discussion forum sentiment is an effective crypto *day trading* strategy and for what time intervals it is most effective, to ultimately help people make

*pump and dump schemes- scheme that attempts to boost the price of a stock through recommendations based on false, misleading or greatly exaggerated statements, rising the price, and the promoters sell the stock at the higher price.

smarter and more informed decisions when dealing with this type of currency.

2 Related Work

Previous studies involving sentiment analysis for crypto price trends have employed a variety of different predictive methods. Abraham et al. chose to use a 15-day window of Google Trends and tweet volume data in a multiple linear regression model to predict the daily Bitcoin closing price [4]. They provided their results in **Figure 1** below, where the green and red markers give train and test points respectively.

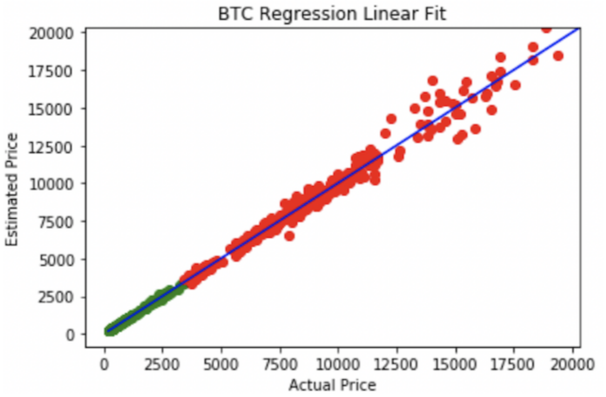

Figure 1. Estimated price as a function of actual price where perfect prediction is given by the blue line.

Another group attempted to make price predictions using both news and social media data, and labeled their data based on actual price changes rather than text sentiment. They also analyzed predicted price fluctuations by percent change and not just direction of movement [5]. Logistic regression proved to be their best model, correctly predicting 43.9% of price increases and 61.9% of price decreases. Included in these correct predictions were all changes of a "large magnitude" relative to the rest (increases greater than 4.9% and decreases greater than 2.71%).

Our project differs from those outlined above in a couple of ways. First, we are examining the correlation between tweet sentiment and Bitcoin price change for much shorter intervals, such as one-hour and six-hours, instead of over a full day. As well, we are deriving our predictive features strictly from the details pertaining to Bitcoin related tweets with no data to Google trends. This relationship is one that has been less frequently studied and could prove to contain valuable insights related to Bitcoin price trends.

3 Methods

3.1 Data Sourcing

The inputs used in our model were retrieved from two sources. A Bitcoin tweets dataset was found on Kaggle and contains 13 features total, with six being relevant to our problem: date (date and time of tweet), user_followers, user_friends, hashtags within the tweet, if the user is verified, and the content of the tweet. Our Bitcoin hourly price data came from Crypto Data Download, having nine total features and three important ones: date (date and time), open and close (coin prices at start and end of the hour respectively). The twitter library spans from February to November of this year, but it is worth noting that there are gaps with days with no tweets at all.

3.2 Data Preprocessing

Data processing occurred in three key ways: cleaning, sentiment scoring, and feature creation. Cleaning involved tasks such as removing irrelevant features and rows with missing values, converting features to appropriate data types.

We then rounded times in the both datasets down to the start of an interval. This would help computational efficiency to determine whether a tweet was within the interval sizes, as all tweets within an interval should now have the same hourly time.

The TextBlob sentiment library was used for sentiment scoring. This lexicon and rule-based analysis tool is specifically attuned to social media sentiments and returns a polarity sentiment for a piece of text that falls between 1 and -1 based on how positive/negative its key words are [6]. Sentiment values were computed for individual tweets, strictly setting the sentiment values to 1, 0, and -1. We then averaged the sentiments in the same time interval. Discussed in a question during our final presentation, we also wanted to add a very strong weight to "influencers", as people like Elon Musk have the strongest ability to vary the price in intraday movement. We created a weighted sentiment, based on the z score of the user's follower count, and averaged this weighted sentiment over the interval as well. All features were normalized before running models.

We also computed an "influencer score" for each user from an impact score calculator found online, according to the following equation:

$$HashtagImpact = 1 + HashtagCount \div 20$$

$$Impact = (isVerified \times HashtagImpact \times Friends \div 4)$$

$$Influence = Followers + Impact$$

(3)

Next, we wanted to simulate using moving avengers in sentiment for gauging possible changes in sentiment. Similar to using a stock price's different interval moving averages as a trading strategy, we added the previous interval, previous 7-interval average, and previous 30-interval averages for both weighted and unweighted sentiment values as well. If the interval average wasn't able to be calculated due to indexing, the mean of the sentiment value was implace as a missing value. In addition to this, we added 3 indicator features that determined if any of the previous indicators were greater or less than our current sentiment, labeled *day_change*, *week_change*, and *month_change* for simplicity, which suggests a recent change in overall sentiment. We added these features to simulate whether larger interval moving averages could aid in predicting the price, as this strategy is often used with normal stock trading.

Mentioned in a conversation with our professor, we wanted to add an indicator for movement given the actual day of the month. Although you can purchase cryptocurrency whenever you want, we believe there is a stronger inclination for bigger moves in the market on Fridays, Mondays, beginning, and end of the month. End of months usually cause selloffs or rallies towards specific target prices. Beginning of the month can also cause rallies due to excitement. We added a feature *isRallyTime* which was a 1 or 0 depending on whether the data fell into one of these categories. Although these variables won't change by time interval during the same day, we wanted to add this feature to signify a date variable to our features, which is also a stock trading strategy.

Concluding, we were incredibly rigorous in revising and creating new features that could be indicators of the price change based loosely on current stock trading strategies. Doing so, we added 13 features and two labels, with the key feature being *sentiment* and *weighted sentiment* (average overall sentiment and weighted sentiment of tweets for a specific time interval). We had two labels to distinguish between a prediction vs. classification problem. For classification, we had *direction* (0 or 1 if the closing price minus open price of Bitcoin for a time interval is negative or positive/neutral). For prediction, we used Bitcoin's closing price. Distributions of relevant variables are given in **Figures 2** and **3** below.

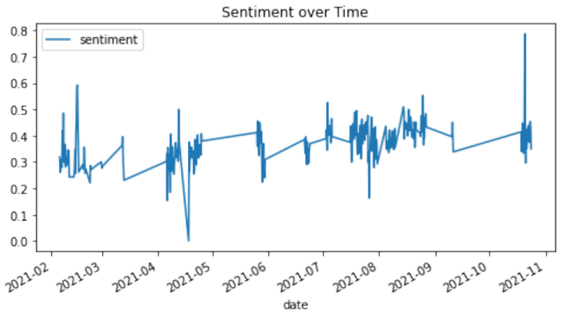

Figure 2. Fluctuation of average ***sentiment*** scores from February to November 2021 for 6-hr time intervals. Note: straight lines represent no twitter data given for that day.

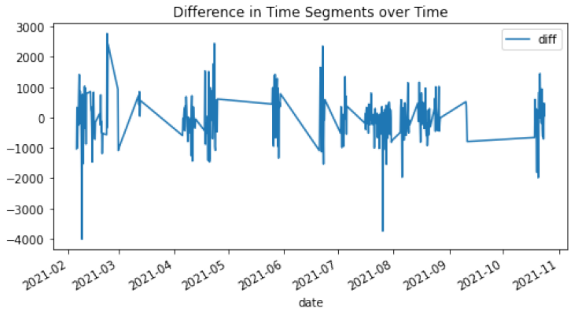

Figure 3. Variation of ***diff*** (close - open Bitcoin price) from February to November 2021 for 6-hr time intervals. Negative diff values translated to a ***direction*** label of 0, while positive or neutral values became a 1.

The aforementioned features were computed for Bitcoin price changes over one, six and twelve hour time intervals. Each time interval was associated with its own dataframe to be used for analysis.

3.3 Feature Selection

Following data processing, sequential forward feature selection (SFS) was run to assess what number of features would result in optimal model performance.

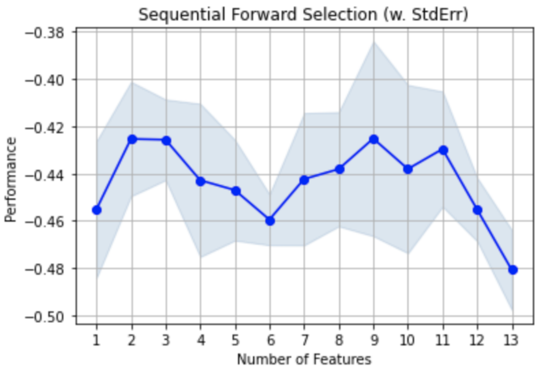

Figure 4. Results of SFS algorithm where the performance of a logistic regression model for the 6-hr time interval was plotted against the number of features used.

Based on the plot, model performance seemed to be most optimized with 9 features, so we decided to remove the last four in order to optimize our performance. We got varied success with this and maintained both feature sets for testing.

3.4 Algorithms

Logistic Regression

The first algorithm that we implemented was logistic regression to analyze the feasibility of predicting Bitcoin price differences as a classification problem. We formulated a case of binary classification, where the two possible outcomes are 0 (decrease in Bitcoin price) and 1 (no change or increase in Bitcoin price) within the respective time interval. The algorithm utilizes a weights vector, W, and features, x, in accordance with **Equation 1** below to output a probability [5]. The label is assigned a 0 or 1 if the probability is below or above 0.5 respectively. **Equation 2** gives the logistic loss function [5].

$$\sigma(x) = \frac{1}{1+\exp(-W^T x)} \tag{1}$$

$$l(x) = \sum_{i=1}^{m} y^{(i)} \log(\sigma(x^{(i)}) + (1-y^{(i)}) \log(1-\sigma(x^{(i)})) \tag{2}$$

Long Short Term Memory (LSTM)

The second algorithm implemented in this project was LSTM. This advanced recurrent neural network (RNN) is sequential in type and allows information to persist [7]. We believe this model would be effective for our time series data, as the model can take into account previous predictions and data. We will be feeding the closing price as the label to test the ability to predict the price. **Figure 5** below gives a high-level overview of this model.

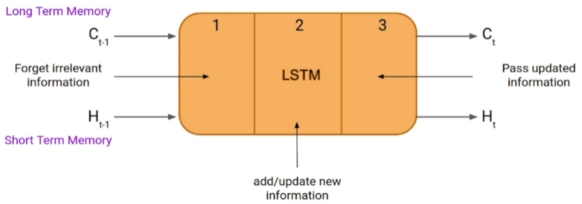

Fig. 5. The three parts and two states of LSTM: forget gate (1), input gate (2), output gate (3), cell state (C) and hidden state (H). As information moves through the model, it is analyzed for relevance, remembered and passed on [7].

4 Results

LSTM

Given that it is an algorithm suited for time series data, the LSTM model was used to predict interval-end Bitcoin prices from historical data. **Figures 6**, **7** and **8** give the achieved results on the validation set.

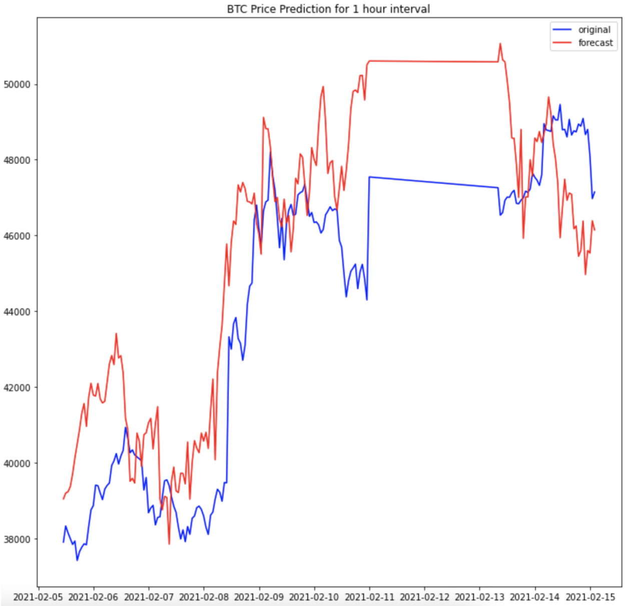

Figure. 6. Original (blue) and predicted (red) bitcoin prices for 1-hr time interval. The R^2 and MSE values were 0.63 and 0.06 respectively.

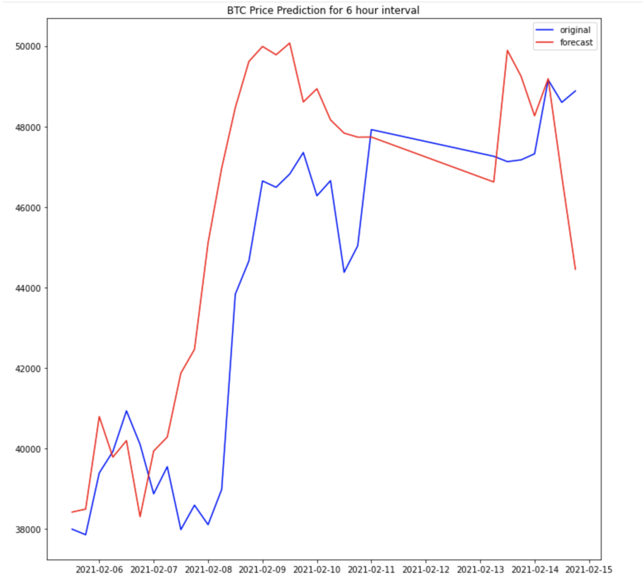

Figure 7. Original (blue) and predicted (red) bitcoin prices for 6-hr time interval. The R^2 and MSE values were 0.38 and 0.09 respectively.

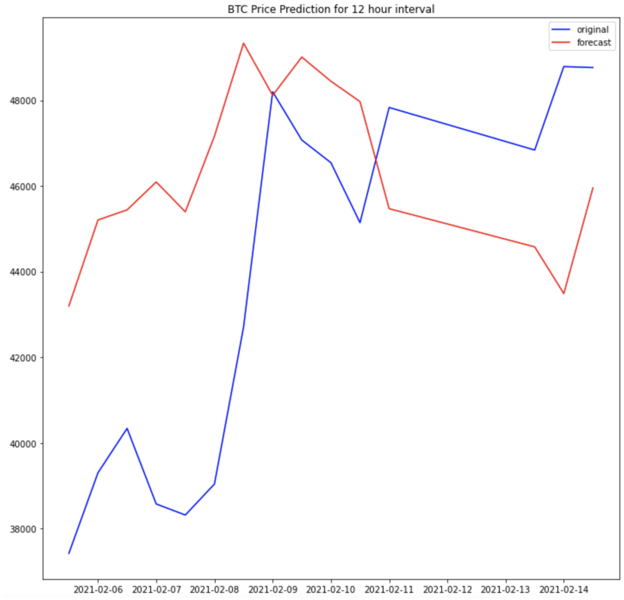

Figure 8. Original (blue) and predicted (red) bitcoin prices for 12-hr time interval. The R^2 and MSE values were -0.39 and 0.23 respectively.

Logistic Regression

As a commonly used classification algorithm, we used a logistic regression model along with the direction label to see if the chosen features were suitable for predicting Bitcoin price movement. Results can be seen in **Figures 9** and **10**.

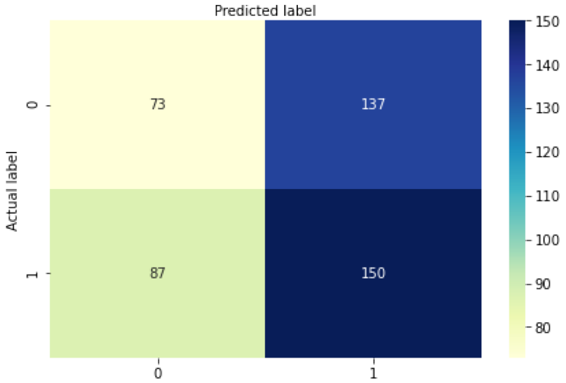

Figure 9. Confusion matrix of logistic regression predictions for 1-hr time interval. The accuracy, precision and recall scores were 0.499, 0.523 and 0.633 respectively.

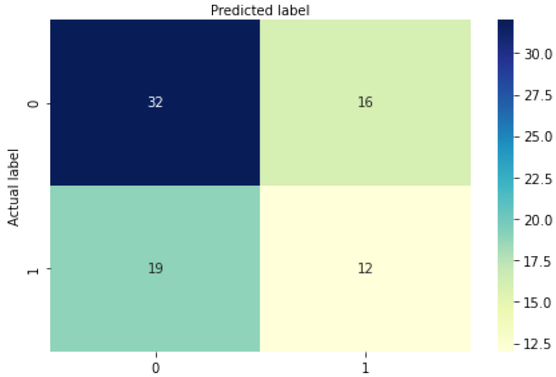

Figure 10. Confusion matrix of logistic regression predictions for 6-hr time interval. The accuracy, precision and recall scores were 0.557, 0.429 and 0.387 respectively.

Confusion matrix of 12-hour not shown for lack of data points: The accuracy, precision and recall scores were 0.357, 0.391 and 0.409 respectively.

5 Discussion

Upon analysis of the results produced by the LSTM and logistic regression models, there are a few points to consider. First, the R^2 value of -0.39 obtained from the model in **Figure 8** is practically impossible. Typically, R^2 values fall between zero and one and assess how well the model fits the actual data. Since the LSTM 12-hr interval model gave a statistic outside of the acceptable range, it is a poor predictor of the actual Bitcoin prices and gives results of questionable validity. We believe this is due to the lack of data points with a 12 hour interval set.

Next, the 1-hr LSTM model seems to have predicted with fairly good accuracy, seen through R^2 and MSE scores close to one and zero respectively. This time interval saw the best performance of the three for the LSTM model, likely because it was run on the largest number of observations. Although the model was run on the validation set for the graphs above, we believe that there is likely overfitting.

Finally, for the logistic regression models, classification was the most successful for the 6-hour interval, with an accuracy of 0.557 as seen in **Figure 10**. While this is not amazing, the model does predict at a rate greater than 50% and the precision and recall scores indicate that the predictions were not just random guesses. It is possible that there was enough contribution from both positive and negative sentiment values for this time interval to most effectively train and test the model, thus generating the optimal results of the group. The confusion matrices do show signs of underfitting, due to that a lot of the sentiment values were positive or neutral. As with LSTM, logistic regression performed the worst for the 12-hour interval, making accurate predictions only 35% of the time. This also can be attributed to the lack of data available for this period.

The relatively average results of our models could be explained by the fact that sentiment analysis, in general, is not as effective for shorter time intervals. It did not show the variation we were looking for and also takes time for the price to reflect the sentiment shown in the predictions, also known as a lagging indicator. As well, sentence level sentiment scoring does not work well with tweets due to their length and uniqueness of language [4]. When looking specifically at Twitter, crypto tweets tend to have mostly "positive" or "neutral" sentiment regardless of the actual price changes occurring. We switched sentiment analysis libraries to try and change this problem. By using a library instead of training a model for sentiment analysis, we are at the mercy of these sentiment values. Twitter activity can also be artificially driven by users with special interests in specific coins, even if they have little knowledge of a coin's behaviour [5].

To improve the performance of our models, we would need a larger dataset in terms of volume and features and a sentiment library more specific to Twitter language. Having 314 rows for 6-hour, and less for 12, isn't enough to form any conclusions. Having stretches of days with no tweets certainly hurt our performance. In the initial data selection process, choosing features from multiple different datasets is known to decrease feature bias and increase prediction accuracy. Also, using a custom made sentiment library would yield more appropriate sentiment scores for context-specific text and filter out some of the neutrality, making it easier to predict price drops.

6 Conclusion

Overall, predicting changes in Bitcoin price solely from sentiment information has proven to be quite difficult. While there is noticeable similarity between the predicted and actual Bitcoin prices for the 1-hour LSTM model (**Figure 6**), from our results as a whole we cannot make firm assertions about the relationship between tweet sentiment and Bitcoin price for time intervals within a day. In reality, there is a strong correlation between social media sentiment and crypto trading decisions for longer time periods (daily+), so our models could be refined to reflect this fact. We could increase our dataset by collecting twitter data over a longer and more continuous amount of time. We could also include features from more sources such as Google Trends or other social media platforms and use a sentiment library more in tune with social media stock context. Ultimately, we are very pleased with trying to solve an unorthodox problem for our final project, where datasets, resources, and "reference" solutions are not as readily available as some other project ideas we could have chosen. We learned firsthand how datasets can be limiting, in both size and feature-set, and steps to resolve real data science problems.

7 Contributions

This project was a fully collaborative effort. AJ's main focus was the coding component while Ziyana focused on synthesizing the information in the presentation and paper. See Github repo, with documented code here.

REFERENCES

[1] Steven Kovach. 2021. *Tesla buys $1.5 billion in bitcoin, plans to accept it as payment.* (February 2021). Retrieved December 12, 2021 from https://www.cnbc.com/2021/02/08/tesla-buys-1point5-billion-in-bitcoin.html

[2] Jemima Kelly. 2021. *Is there a correlation between US Covid cases and crypto prices?*. (August 2021). Retrieved December 5, 2021 from https://www.ft.com/content/aef6cba4-1c94-4b9a-a4b3-329077b72dea

[3] Jack Caporal. 2021. *Tim Cook Owns Cryptocurrency -- So Do 68% Of American Millionaires*. (November 2021). Retrieved December 12, 2021 from https://www.fool.com/research/american-millionaire-crypto-investors/

[4] Abraham, Jethin; Higdon, Daniel; Nelson, John; and Ibarra, Juan (2018) *Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis*, SMU Data Science Review: Vol. 1 : No. 3 , Article 1.

[5] Lamon, C., Nielsen, E., Redondo, E.: *Cryptocurrency price prediction using news and social media sentiment*. Master's thesis, Stanford (2015).

[6] Shubham Jain. 2018. *Natural Language Processing for Beginners: Using TextBlob*. (February 2021). Retrieved December 5, 2021 from https://www.analyticsvidhya.com/blog/2018/02/natural-language-processing-for-beginners-using-textblob/

[7] Shipra Saxena. 2021. *Introduction to Long Short Term Memory (LSTM)*. (March 2021). Retrieved December 5, 2021 from

https://www.analyticsvidhya.com/blog/2021/03/introduction-to-long-short-term-memory-lstm/