The Netherlands Press

Journal of Emerging Strategies in New Economics

Article

STOCK PRICE PREDICTION FROM NEWS HEADLINES USING MACHINE LEARNING MODELS

Ahmed J. Obaid

Asst. Professor, Faculty of Computer Science and Mathematics,
University of Kufa,
Najaf, Iraq

Orchid Id: https://orcid.org/0000-0003-0376-5546

E-mail: ahmedj.aljanaby@uokufa.edu.iq

Abstract.

It's a tremendously intriguing and exciting issue to forecast and speculate on stock market values, especially worldwide company values. This article uses economic news received from businesses to discuss changes in stock prices and projections of stock values. Pay attention to business news headlines and assess headline sentiment using a number of tactics. The Neural Network reconstructs sentiment outcomes with changes in equities over the same period by using BERT as a benchmark and comparing the findings with three other tools. Compared to the other two tools, BERT and RNN are substantially more accurate since they can recognise emotional values without the neutral component. Establish when changes in stock values occur by contrasting these findings with stock value fluctuations over the same time period using sentiment analysis of economic news articles. The impact of sentimental value on changes in sentimental stock market value was also shown to vary significantly amongst the various models.

Keywords: BERT, RNN, Sentiment analysis, Tokenization.

Journal of Emerging Strategies in New Economics Volume 1 Issue 1

Received Date: 08 July 2022
Accepted Date: 11 August 2022
Published Date: 07 September 2022

1. Introduction

Creating and/or using a model that anticipates emotion by identifying word connections and assigning positive or negative sentiment to them is a typical objective. Today, there are numerous options to perform sentiment analysis, including external services like Text Blob that are nearly entirely prepared to utilise it in a particular environment that requires it. Additionally, there are choices that let us build our own models and train them using our own data. One of the most effective methods for sentiment analysis is BERT, however the recurrent neural network (RNN) and the NLTK tool may also be utilised with the sentiment intensity analyser [1].

One of the most significant economic elements is the stock market. Many individuals attempt to characterise and explain various stock market moves in a variety of ways. In spite of the fact that a variety of techniques have been utilised for sentiment analysis, economic news has received special attention, with an emphasis on only the headlines. The headlines of various publications play an even more significant function than ever in the media and news consumption of today [2]. In order to ascertain the effect of headlines on the stock market, sentiment analysis is now used to these headlines concerning a certain firm or companies. If there is a meaningful effect, it is unclear what impact economic headlines have when taken out of the context of economic news.

Analytics is built on a strong foundation of data. It mostly uses economic news headlines for sentiment analysis. Depending on the firm, secondary and other stock market data are also required [3]. There are several options for gathering data and analysing it, from "conventional" human-run dictionaries to "more serious" neural networks that can identify the emotional polarity of each business story's title and assign the appropriate label. There are several techniques accessible to obtain stock market information, including maybe company-specific information, which is extremely significant to us. Work with the most recent information available, based on data given by firms, in any scenario.

Data on stock prices and economic news headlines are linked to the time period that the news covers. The outcomes of the provided sentiment analysis and the variety of stock market data will thus be acceptable. The following parts can be separated from the analysis. gathers stock market data based on the timestamps of specific business news headlines and gathers business news headlines based on firms [8]. After that, prepare the data and use various analytical tools, like as RNN or NLTK. TensorFlow's libraries and functions are used to create and instruct the RNN model. Utilize visualisation and interpretation to manage this data and compare it to sentiment and stock market data. demonstrates how significant economic news stories may have an impact on a variety of developments in the stock market and the general public.

2. NATURAL LANGUAGE TOKENIZATION:

A framework called the Natural Language Toolkit (NLTK) is used to develop Python applications for statistical natural language processing that use data from human language (NLT). It includes text processing libraries for semantic reasoning, root creation, encoding, parsing, categorization, and markup. A common Python package called NLTK offers a wide range of NLP methods. It is one of the most popular libraries for computational linguistics and natural language processing. For Python source code, the encode module offers a lexical scanner that is Python-based. The scanner in this module also provides a token as a response, making it suitable for "beautiful printer" solutions like a colour corrector for on-screen display [7].

When working with textual data, one of the most typical jobs is tokenization. But what exactly does the phrase "tokenization" mean? Tokenization is the process of breaking down a phrase, sentence, paragraph, or complete text document into smaller parts like individual words or keywords.

Each of these smaller components is referred to as a token. Make a list of any phrases that come to mind and keep them in mind while you read this section. It will make it lot easier for you to comprehend the significance of coding.

Before processing a natural language, the words that comprise a character string must be identified. As a result, coding is the first step in implementing NLP (text data). This is significant because studying the words in the text may quickly explain the meaning of the text [4].

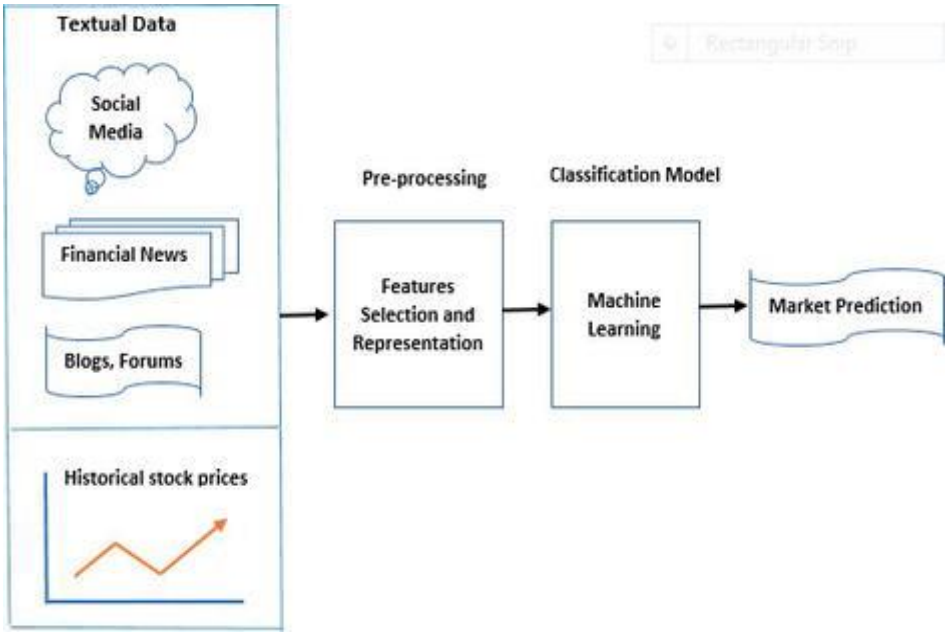

Figure 1: Flow process of stock price prediction[5]

3. IMPLEMENTATION OF STOCK PRICE PREDICTION:

By developing a model that uses Headline News to assess stock prices. There are several sorts of articles, and stock prices change as a result of this study of news headlines utilising sentiment analysis using NLP and projecting whether activity would rise or decrease. This is an example of stock market sentiment analysis. The Kaggle dataset is used in this case. You may get it straight from Kaggle or through colab if the workspace is Google colab.

This dataset is a mix of Kaggle's world news and stock prices. In the dataframe, there are key news title columns for each day, as well as date and label columns (feature dependent). Yahoo Finance has deleted the data for 2020. The Dow Jones Industrial Average stock index is used to create the labels.

	Company	Date	Time	News Headline
0	AMD	2020/11/17	06:30AM	Intel May Be Down, But It's Far From Being Out
1	AMD	2020/11/16	08:02PM	Samsung Intensifies Chip Wars With Bet It Can ...
2	AMD	2020/11/16	04:15PM	AMD to Present at Credit Suisse 24th Annual Te...
3	AMD	2020/11/16	10:29AM	Microsoft (MSFT) Detects Cyberattacks on COVID...
4	AMD	2020/11/16	09:45AM	NVIDIA Looks to Erase a Key AMD Advantage

Figure 2: dataset head[6]

3.1. DATA PRE-PROCESSING:

To retrieve the object label column, remove everything but the object and the date column, and use the label column as the label column. Once the model is developed, they are sent to the network for training.

Data cleaning entails deleting special characters such as "@,!, %," and so on. The % accuracy will drop if these characters are included in the data.

Data in uppercase is changed to lowercase. This data manipulation boosts the model's performance. The training and testing data sets are created by shuffling and filling the data. A cyclic neural network model is used to relay this data.

```
padded_shapes = ([None], ())
train_dataSet = train_dataSet.shuffle(BUFFER_SIZE).padded_batch(BATCH_SIZE, padded_shapes=padded_shapes)
test_dataSet = test_dataSet.padded_batch(BATCH_SIZE, padded_shapes=padded_shapes)
```

```
    model = tf.keras.Sequential([tf.keras.layers.Embedding(encoder.vocab_size, 64),
                                 tf.keras.layers.Bidirectional(tf.keras.layers.LSTM(64, return_sequences=True)),
                                 tf.keras.layers.Bidirectional(tf.keras.layers.LSTM(32)),
                                 tf.keras.layers.Dense(64, activation='relu'),
                                 tf.keras.layers.Dropout(0.5),
                                 tf.keras.layers.Dense(1, activation='sigmoid')])
```

4. BUILDING A RECURRENT NEURAL NETWORK:

The cyclic neural network model is constructed using a sequential model with lexical encoding of input going through a word size of 64, resulting in a word integration layer. Following the integration layer of size 64, a two-way LSTM layer is added. The sequential model is then filled with a two-way LSTM layer of size 32. The sequential model was then filled with a thick layer of 64 neurons with ReLU activation functions. Then, after passing an inhibitory layer by deactivating 50% of the neurons in the hidden layer, the final output layer is placed in a sequential model with a single functional neuron. Because the Sigmoid is at the conclusion of the sequence, it gets activated. model.

4.1. TRAINING THE MODEL:

It is required to optimise and measure in order to train the loss model. The loss function in this case is called binary cross entropy, and it is employed in binary classification tasks. These are assignments that have only two possibilities for answering a question (yes or no, A or B, 0 or 1, left or right). Adam, an optimization technique that substitutes stochastic gradient descent for training deep learning models, is regarded the optimizer in the training model. Adam combines the finest features of the AdaGrad and RMSProp algorithms to provide an optimization technique that can handle sparse gradients in noise issues.

The learning rate, on the other hand, is low and deemed 0.0001, which is ideal for model training. The data is used to assess the model's accuracy. Accuracy is a statistic that describes a model's overall performance across all classes. When all levels are equally significant, it is beneficial. It is derived by dividing the number of right predictions by the total number of guesses.

```
    model.compile(loss='binary_crossentropy',
                  optimizer=tf.keras.optimizers.Adam(1e-4),
                  metrics=['accuracy'])
```

```
fit_model = model.fit(train_dataSet, epochs=5, validation_data=test_dataSet, validation_steps=30)
```

4.2. USING BERT MODEL

Emotion analysis is a significant task in natural language processing (NLP). It is used to determine if customers/people are favourable, negative, or neutral about items, movies, and other comparable things. Among the several advantages that BERT gives to AI are: Model performance is far superior to legacy techniques. Capability to handle more text and languages. A simple method for employing pre-trained models (forward learning)

```
model = text.text_classifier(name='bert', train_data=(X_train, y_train),
                             preproc=preprocess)
```

```
Is Multi-Label? False
maxlen is 400
done.
```

```
learner = ktrain.get_learner(model=model,
                             train_data=(X_train, y_train),
                             val_data=(X_test, y_test),
                             batch_size=6)
```

```
learner.fit_onecycle(lr=2e-5, epochs=1)
```

5. RESULT AND DISCUSSIONS:

The accuracy acquired for the first training epoch was 61.47 percent, and the validation accuracy obtained was 80.78 percent. For the final epoch, the training accuracy gained was 94 percent, and the validation accuracy obtained was 85.73 percent.

```
Epoch 1/5
391/391 [==============================] - 76s 194ms/step - loss: 0.6297 - accuracy: 0.6147 - val_loss: 0.4476 - val_accuracy: 0.8078
Epoch 2/5
391/391 [==============================] - 76s 194ms/step - loss: 0.3471 - accuracy: 0.8646 - val_loss: 0.3556 - val_accuracy: 0.8542
Epoch 3/5
391/391 [==============================] - 76s 194ms/step - loss: 0.2585 - accuracy: 0.9081 - val_loss: 0.3275 - val_accuracy: 0.8708
Epoch 4/5
391/391 [==============================] - 75s 192ms/step - loss: 0.2095 - accuracy: 0.9316 - val_loss: 0.3728 - val_accuracy: 0.8521
Epoch 5/5
391/391 [==============================] - 75s 192ms/step - loss: 0.1859 - accuracy: 0.9402 - val_loss: 0.3559 - val_accuracy: 0.8573
```

For sentiment declaration, if the sentiment acquired after the output layer is equal to 0.5, the sentiment is neutral; if the sentiment is more than 0.5 but less than 1, the sentiment is positive; and if the sentiment ranges between 0 and 0.5, the sentiment is negative. This categorization calculates and analyses the gathered attitudes for various firms.

```
def sentiment_declareation(sentiment_value):
  if sentiment_value == 0.5:
      return 'neu'
  elif 0.5 < sentiment_value <= 1.0:
      return 'pos'
  elif 0.0 < sentiment_value <= 0.5:
      return 'neg'
```

	Company	Date	Time	News Headline	Compound	Sentiment	MaxDate	MinDate
0	AMD	2020-11-17	06:30AM	Intel May Be Down, But It's Far From Being Out	0.419827	neg	2020-11-17	2020-10-27
1	AMD	2020-11-16	08:02PM	Samsung Intensifies Chip Wars With Bet It Can ...	0.730233	pos	2020-11-17	2020-10-27
2	AMD	2020-11-16	04:15PM	AMD to Present at Credit Suisse 24th Annual Te...	0.119468	neg	2020-11-17	2020-10-27
3	AMD	2020-11-16	10:29AM	Microsoft (MSFT) Detects Cyberattacks on COVID...	0.765284	pos	2020-11-17	2020-10-27
4	AMD	2020-11-16	09:45AM	NVIDIA Looks to Erase a Key AMD Advantage	0.245558	neg	2020-11-17	2020-10-27

The positive and negative sentiment is attached and counted for the sentiment analysis computation of data, as indicated in the table below. AMD has the most unfavourable sentiment of any firm in the data. In the statistics provided at the time of 2020, FB has the lowest negative count. When it comes to positive emotion, Facebook has the greatest and AMD has the lowest, and there is no neutral feeling in any of the corporations.

	Company	Neg	Neu	Pos
0	AMD	50	0	49
1	AMZN	39	0	60
2	FB	35	0	64
3	GOOG	45	0	54

The following bar chat displays the favourable and negative attitude of several corporations. Google and Amazon both have positive attitude, although AMD has extremely favourable emotion. In terms of negative feedback, AMAZON and GOOGLE have low levels.

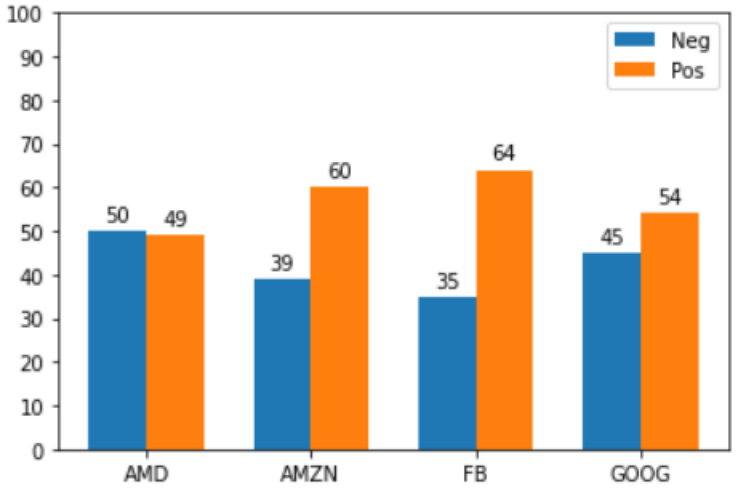

Figure 3: Bar chart of sentiment count of companies

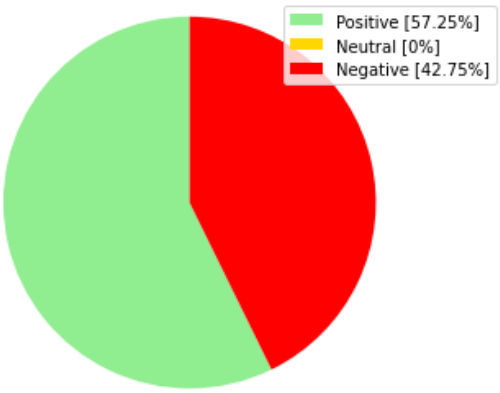

Figure 4: Aggregate sentiment of the economic headlines

The BERT model predicts data and sentiment analysis using the model, and by predicting by testing the data accuracies are also obtained. The BERT model also provides the highest sentiment. Following the data from the RNN model.

```
predictor.predict(texts=testing_data)
```

```
['neg', 'neg', 'neg', 'pos']
```

```
predictor.predict(texts=testing_data, return_proba=True)
```

```
array([[0.997152  , 0.00284804],
       [0.99406946, 0.00593052],
       [0.9803694 , 0.01963067],
       [0.00292435, 0.9970757 ]], dtype=float32)
```

6. CONCLUSION:

Traditional neural networks have outputs and inputs that are completely independent of one another. In the case of a cyclic neural network, however, the output of the previous stages is fed into the current state's input. Recurrent Neural Network - A neural network that is purposely run several times, with bits of each run passed into the next run. The previous run's hidden layers, in particular, offer partial input to the same hidden layer in the next run. Cyclic neural networks are especially good for sequence evaluation because hidden layers may learn from prior neural network runs on earlier sections of the sequence. The picture below, for example, depicts a cyclic neural network looping four times. It's worth noting that the values learnt in the first run's hidden layers are part of the input for the identical hidden layers in the second run. Similarly, the values learnt in the second run's hidden layer become part of the input for the same hidden layer in the third run. The cyclic neural network therefore trains and gradually predicts the meaning of the full sequence rather than the meaning of individual words.

References

[1] H. Alostad and H. Davulcu, "Directional prediction of stock prices using breaking news on twitter," in 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WIIAT), vol. 1, Dec 2015, pp. 523–530.

[2] X. Li, H. Xie, Y. Song, S. Zhu, Q. Li, and F. L. Wang, "Does summarization help stock prediction? a news impact analysis," IEEE Intelligent Systems, vol. 30, no. 3, pp. 26–34, May 2015.

[3] D. Duong, T. Nguyen, and M. Dang, "Stock market prediction using financial news articles on ho chi minh stock exchange," in Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication, ser. IMCOM '16. New York, NY, USA: ACM, 2016, pp. 71:1–71:6. [Online]. Available: http://doi.acm.org/10.1145/2857546.2857619

[4] Y. Wang, D. Seyler, S. K. K. Santu, and C. Zhai, "A study of feature construction for text-based forecasting of time series variables," in Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, ser. CIKM '17. New York, NY, USA: ACM, 2017, pp. 2347–2350. [Online]. Available: http://doi.acm.org/10.1145/3132847.3133109

[5] Y. Shynkevich, T. McGinnity, S. A. Coleman, A. Belatreche, and Y. Li, "Forecasting price movements using technical indicators: Investigating the impact of varying input window length," Neurocomputing, vol. 264, pp. 71 – 88, 2017, machine learning in finance. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231217311074

[6] H. D. Huynh, L. M. Dang, and D. Duong, "A new model for stock price movements prediction using deep neural network," in Proceedings of the Eighth International Symposium on Information and Communication Technology, ser. SoICT 2017. New York, NY, USA: ACM, 2017, pp. 57–62. [Online]. Available: http://doi.acm.org/10.1145/3155133.3155202

[7] R. P. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news: The azfin text system," ACM Trans. Inf. Syst., vol. 27, no. 2, pp. 12:1–12:19, Mar. 2009. [Online]. Available: http://doi.acm.org/10.1145/1462198.1462204

[8] A. de Myttenaere, B. Golden, B. L. Grand, and F. Rossi, "Mean absolute percentage error for regression models," Neurocomputing, vol. 192, pp. 38 – 48, 2016, advances in artificial neural networks, machine learning and computational intelligence. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231216003325