SrockPredictionNLP

by Muhammad Shoaib

Submission date: 10-Aug-2021 03:18AM (UTC+0500)
Submission ID: 1629225051
File name: 103835_Muhammad_Shoaib_SrockPredictionNLP_725148_802192101.pdf (458.16K)
Word count: 2091
Character count: 10749

Stock Market Prediction using NLP over Reddit Data

Abstract:

Reddit has become, since early 2021, a highly popular issue on the financial market. The conversations in these fora suggest that the stock market may have an effect. My project aims to develop a market movement's model based on Reddit's rich text data. In particular, I have studied penalty embedding, CNN-based model embedding, and sentiment analysis methodology to use the statement information to predict the market. This research has evaluated and compared many models. Until now, the performance suggests that the model may enhance the efficiency of the naïve prediction approach marginally.

1 Introduction:

Recently, in financial markets the Reddit forum has been quite popular. It is reported that the subscription WallStreet- Bets is starting a short break on GameStop which leads to a high price hike. In a few days, such a short-term squeeze triggered losses to certain hedge funds of up to USD 70 billion. Reddit has never before been aware of these developments while remaining in the stock market.

I am looking for the answer to the following questions in this project: Does the text of this Reddit provide the information to forecast the march? To this aim, from Jan/2020 to Feb/2021 I have compiled the text of this Reddit and then used the previous trading day test to anticipate the market present day. One of the primary challenges for achieving this assignment is that the text collected is long. The information I have collected amounts to 2.623.978 or 22.868.374 words. In the meantime, we only have some two hundred trading days of information if we create a model with a daily frequency.(Řpiewak et al., 2018)

Due to this imbalance of dependent and separate variables, the key to effective prediction is a model with sufficient complexity. Too many parameters can result in over fitting a model, and information cannot be captured by an over fit model with too few parameters.

1) Data was obtained from 2 Reddit APIs, and pre cleaned my technique may be summarized as follows

2) Raw data are then used to create numerical measurements with a) BERT-based sentence embedded,(Gan et al., 2017)(b) and each Reddit post is merged as a document that embeds with(Bollen et al., 2011). The TextBlob and (Fulda et al., 2017) leverages have been additionally given a daily feeling. 3) A CNN model for the prediction of stock moving men is then created;

2 Literature Review:

Many studies on extracting information from phrases and texts are available. A technique to employ an average of work vectors to construct sentence representations has been proposed by (Reimers & Gurevych, 2020). The paradigm for the encoding of sentences was created. The text continuity is used to create an encoder decoder model by Ryan, Yukun, Ruslan, et al., and then uses the vector produced for the phrase. In sentencing representations, BERT was also employed. Sentence-BERT (Oncharoen & Vateekul, 2018) changes the pre-trained BERT network that employs Siamese and triple network structures to produce semantic-significant phrases in the network. Another Author also propose a vector for the paragraph allowing multiple text lengths, such as document or paragraph, to be embedded.

Financial forecasting with the NLP approach is also very often undertaken. A review of the new articles on the natural language financial prediction was carried out by (Arora et al., 2016) This article examines the studies from the 1980s in this subject and also provides some advancements in the use to predict social media content. Many research focus on utilizing Twitter data in the subject of levying material from social media due to its simple semanticization and limited duration.
In this sector, machine learning methods were also frequently employed. The neural network tensor and convolutionary (Carli, 2003)are used to foresee a news text- utilized the Twitter feeling score for predicting stock market motion. (Cho & Merri, 2013)

In the earlier CS230N studies, the financial forecasting field was also studied by the NLP technique. (Fahmy et al., 2020)the FOMC meeting transcript to forecast the FRB rate changes in sentiments embedding and document embedding. Utilizes the document incorporation approach to forecast the famous variables in the French equity market. Another researcher uses an eight-kg Company report focused model to forecast post-earnings price volatility.
The Reddit occurrence occurs however lately and may be continuing to evolve. There is not much study into constructing NLP models on this information in order to broaden my knowledge which is the objective of my project.

3 Dataset Explanation:

The data set is used from kaggle and the link of the data is https://www.kaggle.com/aaron7sun/stocknews.
Two types of data provided in this dataset:

1. News data: Browsed through Reddit World News Channel's historic news headlines. They are classified with the revised votes of the users, and just a single date is taken in the top 25 headlines.

2. Stock Data: "Proving the idea" is based on the (DJIA).

All data supplied in 3.csv format data files:

1. The column attribute is "date, and “news headlines." RedditNews.csv: 2 columns. Every story is classified in terms of how hot they are from top to end
2. DJIA table.csv: got from Yahoo Finance: see the website for additional information.
3. CombinedNewsDJIA.csv: this combined dataset with different 27 features
4. The first column is "Date," the second column is "Label," and the next column is "Top1" through "Top25.

3.1 Dataset inspection and preparation

Load the dataset and print the data first.

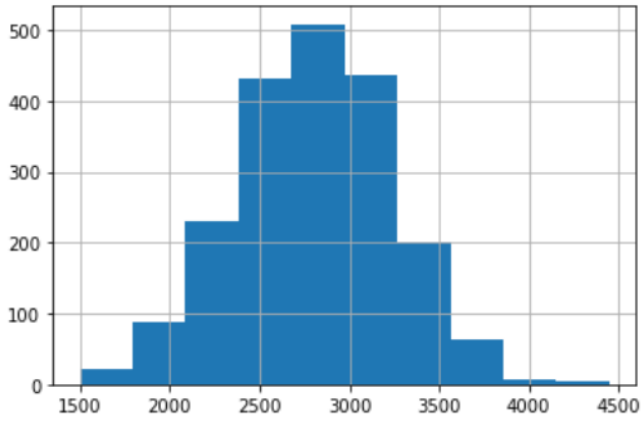
The file consists of the headline Top25 and a label containing just two values:

1. DJIA Adj - "1" Close value increased or stayed identical
2. "0" reduced by the value DJIA Adj Close

Label	Top1	Top2	Top3	Top4	Top5	Top6	Top7	Top8	...
0	b"Georgia 'downs two Russian warplanes' as cou...	b'BREAKING: Musharraf to be impeached.'	b'Russia Today: Columns of troops roll into So...	b'Russian tanks are moving towards the capital...	b"Afghan children raped with 'impunity,' U.N. ...	b'150 Russian tanks have entered South Ossetia...	b"Breaking: Georgia invades South Ossetia, Rus...	b"The 'enemy combatent' trials are nothing but...	...
1	b'Why wont America and Nato help us? If they w...	b'Bush puts foot down on Georgian conflict'	b"Jewish Georgian minister: Thanks to Israeli ...	b'Georgian army flees in disarray as Russians ...	b"Olympic opening ceremony fireworks 'faked'"	b'What were the Mossad with fraudulent New Zea...	b'Russia angered by Israeli military sale to G...	b'An American citizen living in S.Ossetia blam...	...

First, all titles are combined in a single column

Histogram of the headline length



Divide the data across the train and test set. This is about 80%/20% divided.

3.2 Tf-idf and implementation of Simple classification:

The phrases were played into words by using the RegexpTokenizer nltk, starting from the daily headlines. There is also an application of lemmatization. This is the responsibility of the following CustomTokenizer.

The TfidfVectorizer converts the raw headlines into a TF-IDF feature matrix with the CustomTokenizer.

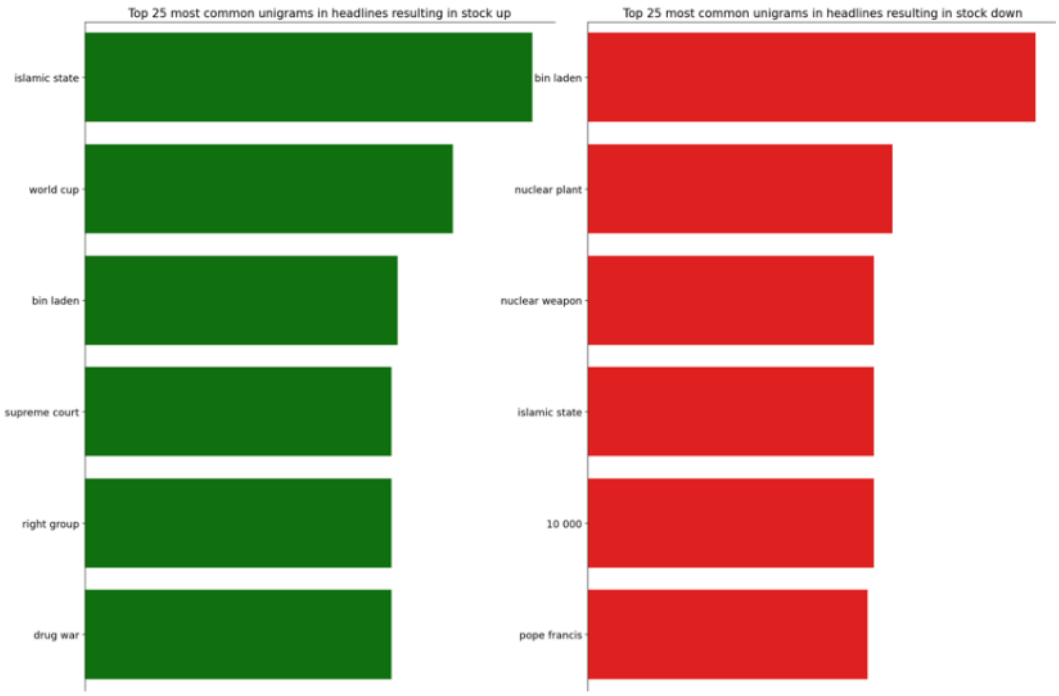
- English stop words are deleted, words are transformed from upper case to lowercase.
- The ngram_range are taken into account
- Min_df Ignore words with strictly higher document frequency than the particular threshold the training dataset used for vectorization.

3.3 Inspect the name and characteristics of the TF-IDF

```
['250 000', '3 000', '3 billion', '3 day', '3 million', '3 year', '30 000', '30 million', '30 year', '300 000', '300 million',
```

	000 000	000 dead	000 euro	000 troop	000 year	1 2	1 3	1 4	1 5	1 500	...	year since	year u	year world	york time	young child	young girl
0	0	0	0	0	0	1	0	0	0	0	...	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0

3.4 Some Common Headlines of the data



3.5 Prediction of Test Data using support vector classifier

```
Accuracy X_train: 0.9621353196772191
```

Predicted	0	1
Actual		
0	36	150
1	30	162

3.6 Confusion Matrix

```
              precision    recall  f1-score   support

           0       0.55      0.19      0.29       186
           1       0.52      0.84      0.64       192

    accuracy                           0.52       378
   macro avg       0.53      0.52      0.46       378
weighted avg       0.53      0.52      0.47       378

Accuracy X_test: 0.5238095238095238
```

4 Implementation of Keras Sequential model:

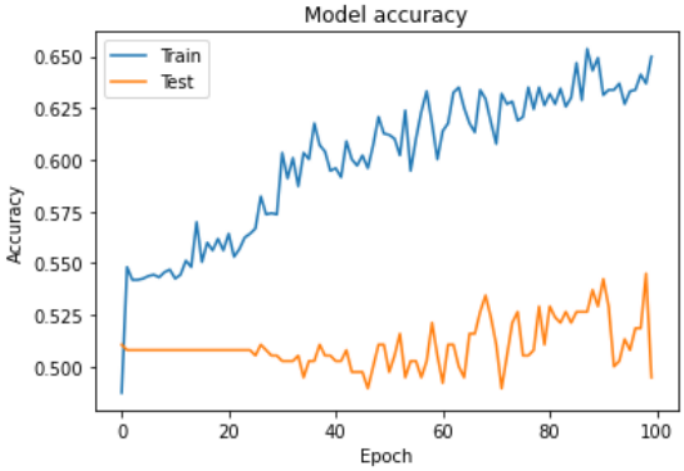
There is a sequence and functionality of two approaches to construct Keras models.
For most situations, the sequential API enables you to construct layer-by-layer models. The fact that you can not construct models which share layers or have many inputs or outputs is restricted. The functional API also gives you the possibility to construct much more flexible models, because you can simply define models whose layers are more connected than only the past and the next levels. You can actually link layers to any other layer (literally). This makes it possible to create complicated networks, such as Siamese and residual networks.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 8)                 9760
_________________________________________________________________
dropout (Dropout)            (None, 8)                 0
_________________________________________________________________
dense_1 (Dense)              (None, 2)                 18
=================================================================
Total params: 9,778
Trainable params: 9,778
Non-trainable params: 0
```

4.1 Model Accuracy:

Model precisions are defined as the number of classifications that are split by the total number of predictions that a model properly forecasts. It is a technique to evaluate the achievement of a model, but definitely not the only way

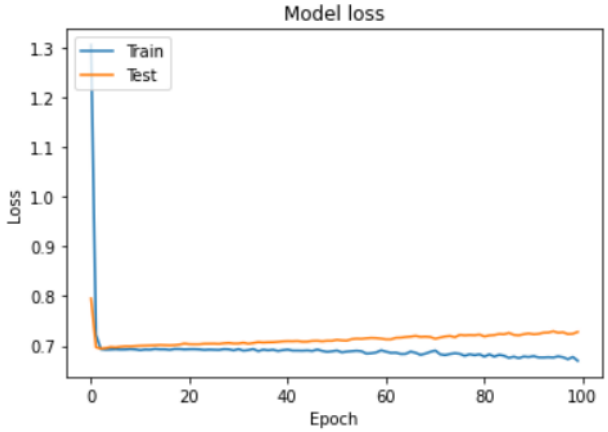
Precision is the number of predictions by the total number of records made in the model. Precision is not a meaningful measure of model performance for an unbalanced data set. For a collection of data with a default rate of 5%, while all records are forecast as 0, the model still has a 95 percent accuracy



4.2 Model Loss:

Loss is the fine for a wrong forecast. That is to say, loss is a number which shows how terrible the forecast of the model was with one sample. If the forecast of the model is perfect, the loss must be zero, otherwise the loss shall be larger. The objective of the model training is to identify,

in all cases, a range of weights and partialities with small losses.




```
12/12 [==============================] - 0s 1ms/step - loss: 0.7282 - accuracy: 0.4947
[0.7281633615493774, 0.49470898509025574]
```

5 Use Word2Vec:

The Word2Vec model extracts the notion of connection through words or products such as semantic connection, synonym identification, categorization of concept, preferences for selecting and analogies. A Word2Vec model learns significant relationships and encodes the connectivity into a vector similarity.

To identify the most related terms, use Word2Vec

```
Number of word vectors: 29909
[('abusing', 0.9775549173355103), ('minor', 0.9760563969612122), ('murdering', 0.972906768321991), ('hanged', 0.9712677001953125),
```

5.1 Some Data Preprocessing used in this method:

5.1.1 Removing Stop words:

Stops words are the words which usually used in any language and in the case of English language there are number of stop words i.e is, are, am, the etc. and these words are not beneficial for training the data in NLP so we can remove these words and the benefit of removing these words is that we have data in more cleaned form.

5.1.2 Tokenization

Tokenization is an important technique in NLP which converts the text into tokens like there are 5 words in a sentence and after tokenization it will be five tokens and it will call word tokenization.

5.1.3 Stemming and Lemmatization:

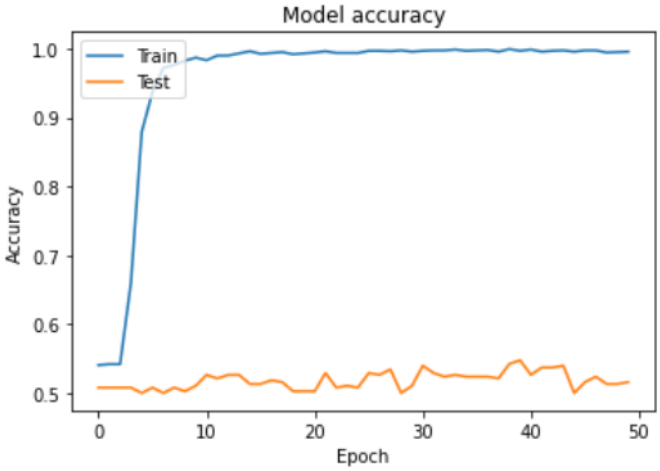
Stemming and lemmatization are procedures used to evaluate the meaning behind a word in search engines and catboats. Stemming employs the word's stem and the context in which the term is employed in the lemmatization. More explanations and examples will be provided later. We want relevant results to be found not just for the precise term that we put into the search field, but also for any alternative versions of the words we use when performing a search.

6 Word embedding

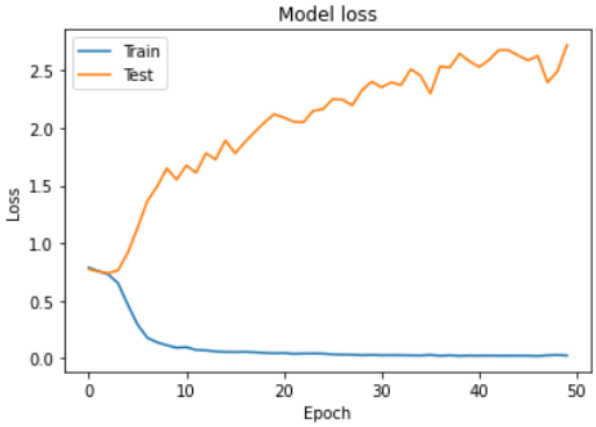
3
An embedding word is a learning representation of a text with a comparable meaning in words. This method is one of the main advances of in-depth study on the hard challenges in the processing of natural languages.

```
Number of words: 20000
max_length: 200
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, 200, 16)           320000
_________________________________________________________________
lstm (LSTM)                  (None, 8)                 800
_________________________________________________________________
dropout_1 (Dropout)          (None, 8)                 0
_________________________________________________________________
dense_2 (Dense)              (None, 16)                144
_________________________________________________________________
dense_3 (Dense)              (None, 2)                 34
=================================================================
Total params: 320,978
Trainable params: 320,978
Non-trainable params: 0
```

6.1 Model accuracy of word embedding Model:



6.2 Model Loss of Word Embedding:




```
12/12 [==============================] - 0s 22ms/step - loss: 2.7192 - accuracy: 0.5159
Testing Accuracy is 51.58730149269104
```

7 Conclusion:

We have predicted stock market price using Reddit data and applied different natural language processing techniques to predict to high and low market prices. The accuracy of the models is slightly low because the models is dependent on news of market data and we don't have time series data or some real numeric data. We can also improve the accuracy of prediction prices

using other Machine learning and deep learning models but I used NLP techniques over them which was basic goal of this course.

Arora, S., Liang, Y., & Ma, T. (2016). A simple but though Baseline for Sentence Embeddings. *Iclr*, *15*, 416–424. e

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

Carli, D. (2003). Honk If You Support JDF. *Print on Demand*, *9*(2), 42.

Cho, K., & Merri, B. Van. (2013). *Learning Phrase Representations using RNN Encoder – Decoder*.

Fahmy, F. K., Khalil, M. I., & Abbas, H. M. (2020). A transfer learning end-to-end arabic text-to-speech (tts) deep architecture. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12294 LNAI*, 266–277. https://doi.org/10.1007/978-3-030-58309-5_22

Fulda, N., Ricks, D., Murdoch, B., & Wingate, D. (2017). What can you do with a rock? Affordance extraction viaword embeddings. *IJCAI International Joint Conference on Artificial Intelligence*, *0*, 1039–1045. https://doi.org/10.24963/ijcai.2017/144

Gan, Z., Pu, Y., Henao, R., Li, C., He, X., & Carin, L. (2017). Learning generic sentence representations using convolutional neural networks. *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2390–2400. https://doi.org/10.18653/v1/d17-1254

Oncharoen, P., & Vateekul, P. (2018). Deep learning using risk-reward function for stock market prediction. *ACM International Conference Proceeding Series*, 556–561. https://doi.org/10.1145/3297156.3297173

Reimers, N., & Gurevych, I. (2020). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 3982–3992. https://doi.org/10.18653/v1/d19-1410

Řpiewak, M., Sobecki, P., & Karaś, D. (2018). *OPI-JSA at SemEval-2017 Task 1: Application of Ensemble learning for computing semantic textual similarity*. 139–143. https://doi.org/10.18653/v1/s17-2018

SrockPredictionNLP

ORIGINALITY REPORT

2%	0%	1%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to North South University Student Paper	1%
2	Xiaoya Yin, Wu Zhang, Wenhao Zhu, Shuang Liu, Tengjun Yao. "Improving Sentence Representations via Component Focusing", Applied Sciences, 2020 Publication	1%
3	"Artificial Neural Networks in Pattern Recognition", Springer Science and Business Media LLC, 2020 Publication	<1%
4	dokumen.pub Internet Source	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		