Tayfun Ayazma, Ph.D.

tayfunayazma@gmail.com

940-595-8413

THE USE OF NLP AND DEEP LEARNING ON NEWS HEADLINES TO PREDICT STOCK MOVEMENTS

May 20, 2020

Table of Contents

Abstract

This project attempts to predict the stock movements of Dow Jones Industrial Average (DJIA) using news headlines scraped from the Reddit WorldNews Channel (r/worldnews). In particular, this project uses the textual data of historical news headlines to predict whether the DJIA index will go up or down. Various advanced machine learning algorithms were experimented along with several deep learning neural network architectures including a fully-connected multilayer perceptron (MLP), recurrent neural network (RNN), and convolutional neural network (CNN) for the predictions. Among others, the recurrent neural network – LSTM architecture achieved the highest accuracy score with a nearly 58% on the validation dataset.

Introduction

A significant number of companies in the financial services and banking industry have long spent millions of dollars and dedicated a vast amount of resources to quantify and analyze qualitative data from a variety of media outlets in order to gain valuable information in making investment decisions. Despite its potential, there is limited success in stock market prediction due to the wide range of difficulties which lie in the complexities of modeling stock market dynamics (Schumaker and Chen, 2009). Yet, in recent years, with the advancement in technology, especially in the area of artificial intelligence, the use of textual data from a variety of resources including SEC reports such as 8-K and 10-K (Lee et al., 2014), financial news articles such as Financial Times and Wall Street Journal (Schumaker and Maida, 2018) and the media outlets such as Twitter, Facebook, and Reddit (Pagolu et al., 2016, Oliveira, Cortez, and Areal, 2018) has gained a significant importance and provided very promising results in stock price prediction. Hence, thousands for documents from different sources have been scraped by investors to extract meaningful information through various NLP techniques to predict stock prices.

In this project, I attempt to demonstrate the viability of the use of natural language processing techniques to extract information from the Reddit news headlines to predict stock movements. Prior studies and projects have shown how useful natural language processing word embeddings extracted from the social media outlets such as Twitter and Facebook are in

predicting stock prices. In this respect, this particular project builds upon the past work by using the NLP word embeddings and various deep learning neural network architectures to predict the movements of the Dow Jones Industrial Average (DJIA) index in the U.S. stock market. The DJIA is a stock market index that tracks the stock performance of 30 large, publicly owned companies (blue-chip companies) listed on stock exchange in the United States (See the companies listed in DJIA below).

Table 1: Components of the Dow Jones as of April 6, 2020

Company Name	Ticker	Exchange
3M Company	MMM	NYSE
American Express Company	AXP	NYSE
Apple Inc.	AAPL	NASDAQ
Boeing Company	BA	NYSE
Caterpillar Inc.	CAT	NYSE
Chevron Corporation	CVX	NYSE
Cisco Systems, Inc.	CSCO	NASDAQ
The Coca-Cola Company	KO	NYSE
Dow Chemical Company	DOW	NYSE
Exxon Mobil Corporation	XOM	NYSE
The Goldman Sachs Group, Inc.	GS	NYSE
The Home Depot, Inc.	HD	NYSE
Intel Corporation	INTC	NASDAQ
International Business Machines Corporation	IBM	NYSE
Johnson & Johnson	JNJ	NYSE

Company Name	Ticker	Exchange
JPMorgan Chase & Co.	JPM	NYSE
McDonald's Corporation	MCD	NYSE
Merck & Co., Inc.	MRK	NYSE
Microsoft Corporation	MSFT	NASDAQ
Nike, Inc.	NKE	NYSE
Pfizer Inc.	PFE	NYSE
The Procter & Gamble Company	PG	NYSE
Raytheon Technologies	RTX	NYSE
The Travelers Companies, Inc.	TRV	NYSE
UnitedHealth Group Inc.	UNH	NYSE
Verizon Communications, Inc.	VZ	NYSE
Visa Inc.	V	NYSE
Walgreens Boots Alliance, Inc.	WBA	NASDAQ
Walmart Inc.	WMT	NYSE
The Walt Disney Company	DIS	NYSE

Method

Data

The data for this project were acquired from the Kaggle.com. Two sets of datasets were available for Reddit historical news headlines and Dow Jones Industrial Average (DJIA) stock data. The data for the news headlines were crawled from the Reddit WorldNews channel (r/worldnews). Only the top 25 news headlines were considered for a single date. The DJIA stock data were downloaded directly from the Yahoo Finance. The combined dataset includes a date column and a label column along with 25 news headlines columns. The “Label” column is the target feature and consists of the values of 0 and 1. While 1 indicates that the DJIA adjusted close value stayed

the same or went up on that date, 0 represents that it dropped on that particular date. During the period of this project, 1065 times the stock prices for DJIA either stayed the same or went up, whereas 924 times went down.

Data Preparation and Text Preprocessing

Once the data were loaded and inspected for quality (outliers, missing values, data types), all the textual data were preprocessed by lowercasing, removing stopwords, punctuation, and numbers, and lemmatizing words using the NLTK library. A "text_preprocess()" function was created to be used in the subsequent sections. Also, a function called "combine_text_columns()" was created to turn all text in each row of the dataframe to a single vector. The last step in data preparation was to split data into train and test sets. Per instructions, I used all of the dates up until 12-31-2014 as the training set and the following two years as the test set which corresponds to roughly 80% training set and 20% test set.

Evaluation

Model evaluation is an important component of machine learning. Hence, two separate functions were created in advance to evaluate model performances in the subsequent sections. The "Evaluation()" function was created to generate a classification report, confusion matrix, and ROC-AUC score per a given model as well as several model statistics including accuracy, precision, recall, and f1 scores. The "ROCCurve()" function, on the other hand, was created to compute and plot ROC curve and AUC score for a given model.

Model Building

Six advanced supervised machine learning models including Logistic regression, Multinomial Naïve Bayes, Random Forest Classifier, Support Vector Machine Classifier, XGBoost Classifier, and Bernoulli Naïve Bayes were constructed in a pipeline using Scikit Learn's Pipeline module. In addition to these models, three deep learning neural network architectures including a fully-connected multilayer perceptron (MLP), recurrent neural network (RNN), and convolutional neural network (CNN) were constructed using Keras with a Tensorflow backend. Each network was trained for 10 epochs.

For the Logistic regression model, a dataframe which consists of the most important features/words both positive and negative was produced. Although the top 10 most important

words with the highest positive coefficients do not seem to be interesting, those negative ones such as "military", "hacking", "criminal", "low", and "sanction" make sense as they have a negative impact on the DJIA stock movement.

Table 2: The Most Important Features

	Word	Coefficient		Word	Coefficient
6042	nigeria	0.838646	4201	hour	-0.684081
4705	jew	0.713932	5633	military	-0.702586
7247	record	0.664518	3852	hacking	-0.703074
8131	since	0.656317	9798	without	-0.739373
8474	state	0.654680	2014	criminal	-0.761785
7556	right	0.653375	5259	low	-0.770794
5825	mubarak	0.640013	7750	sanction	-0.794457
7963	set	0.621422	1970	country	-0.804450
8857	tear	0.617964	4595	iran	-0.811626
7852	scrap	0.600945	7670	run	-0.820194

Model Tuning

The SVM Classifier and the RNN – LSTM network achieved the highest accuracy and AUC scores on the validation dataset among the other models. In order to improve the models' performance, tuning was performed for the some of the important hyper-parameters of the models using the GridSearchCV. However, it did not make a big difference on the performance of the both models.

Table 3: Models with Performance Statistics

Model	Accuracy	ROC_AUC	Precision	Recall	F1
LSTM	0.582011	0.581989	0.589474	0.583333	0.586387
SVM Classifier	0.537037	0.533350	0.530686	0.765625	0.626866
MLP	0.515873	0.514701	0.520737	0.588542	0.552567
Logistic Regression	0.515873	0.509241	0.512968	0.927083	0.660482
XGBoost	0.510582	0.508233	0.514286	0.656250	0.576659
Multinomial Naive Bayes	0.507937	0.500000	0.507937	1.000000	0.673684
BernoulliNB	0.502646	0.495800	0.505682	0.927083	0.654412
CNN	0.497354	0.494792	0.504000	0.656250	0.570136
Random Forest Classifier	0.484127	0.479503	0.494983	0.770833	0.602851

Results

The RNN – LSTM network was able to achieve an accuracy of 58% and AUC score of 0.58 on the validation data. As some claim, deep learning neural networks perform better on big data. Adding more textual data or using pretrained word embeddings could yield a better accuracy score.

Business Recommendations

1. Vast amount of resources is being dedicated every year to make better investment decisions in the financial and banking industry. As this project indicates, the use of NLP on the textual data extracted from different sources is very promising in providing better informed decisions. Investment banks, corporate finance offices and anyone else who is interested or involved in trading securities on public markets as the clients of this project can take advantage of the qualitative data in making better investment decisions.
2. Since the advent of big data and powerful GPUs, the deep learning neural networks have the potential to revolutionize the stock market trading.
3. Changes in stock movements was only measured immediately after news headlines release. However, it is also quite possible that stock movements may react in the days following the release of news headlines.

Conclusion

In this project, I attempted to predict the DJIA stock movements using the historical news headlines extracted from the Reddit. The highest accuracy score that I obtained from the models was 58% on the validation dataset. There are many factors affecting the accuracy score for stock predictions. First of all, I had limited data which affects the performance of the models. Therefore, adding more data would increase the performance of the models in predicting the DJIA stock movements. Also, the stock prediction itself is a very difficult task due to the wide range of difficulties which lie in the complexities of modeling stock market dynamics. Considering the difficulty of predicting stock movements and under these limitations, overall,

our models did pretty good in predicting the stock movements of the Dow Jones. Although this project only touches the surface of NLP techniques here, a 58% accuracy score suggests that the efforts to extract textual data from a variety of sources to make investment decisions could be worth the while.

Citations

Lee, H., Surdeanu, M., MacCartney, B., & Jurafsky, D. (2014). On the Importance of Text Analysis for Stock Price Prediction. In *LREC* (Vol. 2014, pp. 1170-1175).

Oliveira, N., Cortez, P., & Areal, N. (2017). The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with Applications*, *73*, 125-144.

Pagolu, V. S., Reddy, K. N., Panda, G., & Majhi, B. (2016). Sentiment analysis of Twitter data for predicting stock market movements. In *2016 international conference on signal processing, communication, power and embedded system (SCOPES)* (pp. 1345-1350).

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems (TOIS)*, *27*(2), 1-19.

Schumaker, R. P., & Maida, N. (2018). Analysis of Stock Price Movement Following Financial News Article Release. *Communications of the IIMA*, *16*(1), 1.