Game industry companies stock price prediction using GAN and BERT

Jakub Wujec

April 26, 2022

1 Data

1.1 Stock Data

In our analysis we decided to analyze the algorithm on 4 large companies from the gaming industry. This sector was chosen due to the fact that the study takes into account the influence of sentiment analysis - one of our hypotheses is that the opinions of Reddit users have a particularly significant impact on valuations of a given company in the gaming industry. As many as 4 companies were chosen for the analysis - Electronic Arts, Ubisoft, Take-Two Interactive Software, Activision Blizzard, called in the later part of the paper by their stock ticker: EA, UBSFY, TTWO, ATVI. The selection of more than one company was influenced by making sure that the algorithm was reproducible and universal. Furthermore, attempts were made to use the model learned on one company's data on another company's test data. These companies are among the leaders in the growth industry. The choice of only U.S. companies was due to a possible language barrier when analyzing the sentiment of companies that produce games primarily for the Asian market. Due to the use of one-day intervals during replacement, limitations of Reddit's API (Application Programming Interface), and limitations of computing power, the analysis was conducted over a nearly three-year period: from 01/01/2019 to 31/10/2021. The data comes from the US NASDAQ stock market and was retrieved using the yahoo finance library for the python language.

Table 1: Descriptive statistics for chosen companies

	Company	ATVI	EA	TTWO	UBSFY
Open	count	7.570000e+02	7.570000e+02	7.570000e+02	7.570000e+02
	mean	6.837792e+01	1.174786e+02	1.434314e+02	1.535412e+01
	std	1.811106e+01	2.062370e+01	3.220382e+01	2.450012e+00
	min	3.956548e+01	7.472812e+01	8.471000e+01	9.050000e+00
	25%	5.210339e+01	9.717636e+01	1.172500e+02	1.393000e+01
	50%	6.857000e+01	1.193464e+02	1.383700e+02	1.548000e+01
	75%	8.241098e+01	1.385884e+02	1.711000e+02	1.681500e+01
	max	1.033197e+02	1.485431e+02	2.104765e+02	2.089000e+01
High	count	7.570000e+02	7.570000e+02	7.570000e+02	7.570000e+02
	mean	6.922366e+01	1.189510e+02	1.453735e+02	1.546871e+01
	std	1.817212e+01	2.062690e+01	3.241062e+01	2.460111e+00
	min	4.119361e+01	7.691432e+01	8.757000e+01	9.270000e+00
	25%	5.284490e+01	9.860734e+01	1.190800e+02	1.396000e+01
	50%	7.004000e+01	1.206978e+02	1.407800e+02	1.564000e+01
	75%	8.327678e+01	1.402643e+02	1.728500e+02	1.694000e+01
	max	1.040263e+02	1.495559e+02	2.149100e+02	2.134000e+01
Low	count	7.570000e+02	7.570000e+02	7.570000e+02	7.570000e+02
	mean	6.738189e+01	1.158599e+02	1.413383e+02	1.521634e+01
	std	1.794660e+01	2.052602e+01	3.182764e+01	2.440169e+00
	min	3.908489e+01	7.344622e+01	8.441000e+01	9.050000e+00
	25%	5.142120e+01	9.571559e+01	1.158500e+02	1.374000e+01
	50%	6.707523e+01	1.170509e+02	1.363400e+02	1.538000e+01
	75%	8.128642e+01	1.370253e+02	1.689600e+02	1.667000e+01
	max	1.020559e+02	1.457801e+02	2.094350e+02	2.067000e+01
Close	count	7.570000e+02	7.570000e+02	7.570000e+02	7.570000e+02
	mean	6.832515e+01	1.174449e+02	1.434253e+02	1.533973e+01
	std	1.802681e+01	2.056176e+01	3.208114e+01	2.456598e+00
	min	3.933990e+01	7.425114e+01	8.463000e+01	9.140000e+00
	25%	5.210339e+01	9.698756e+01	1.176000e+02	1.384010e+01
	50%	6.947362e+01	1.190582e+02	1.381900e+02	1.551000e+01
	75%	8.230000e+01	1.384656e+02	1.706700e+02	1.682000e+01
	max	1.033098e+02	1.482325e+02	2.133400e+02	2.124000e+01
Volume	count	7.570000e+02	7.570000e+02	7.570000e+02	7.570000e+02
	mean	7.852192e+06	3.352595e+06	1.682517e+06	1.603439e+05
	std	4.732384e+06	2.796399e+06	1.258285e+06	3.360945e+05
	min	1.562888e+06	6.060640e+05	2.116420e+05	1.188800e+04
	25%	5.243074e+06	2.068342e+06	1.023114e+06	3.837000e+04
	50%	6.692846e+06	2.723601e+06	1.351017e+06	5.767100e+04
	75%	8.966312e+06	3.757614e+06	1.913890e+06	1.117520e+05
	max	5.170888e+07	3.870450e+07	1.894501e+07	3.997347e+06

1.2 Text Data

Due to the inclusion of sentiment analysis in our study, the source of textual data sourcing had to be chosen. Due to the fact that we have chosen in the study only the companies from the growth industry, we decided to use data coming from the reddit.com portal, which is one of the largest forums in the world. It also has a feature that helps us to obtain data for our study - it is divided into parts, the so-called subreddits, which gather people interested in a particular topic. For our study, we chose the r/Games subreddit. This is the largest forum on this site dealing with the subject of games. There are more than 3.1 million users on it. Next, for each company selected by us, keywords were chosen to retrieve the data. The keywords were the names of the most popular game series of a particular publisher and the publisher's name itself. Next, using the psaw library for python, we found all the comments that had been posted over a predefined period of time on the r/Games subreddit containing the keywords mentioned below

Table 2: Key words chosen for each company

	EA	TTWO	UBSFY	ATVI
0	EA	Take Two	Ubisoft	Blizzard
1	Fifa	NBA 2K	Assasin's Creed	Starcraft
2	The Sims	Battleborn	AC	Warcraft
3	Need for Speed	BioShock	Far Cry	Overwatch
4	NFL	Borderlands	Watch Dogs	Diablo
5	Apex	Evolve	Rainbow Six Siege	World of Warcraft
6	Battlefield	Mafia	Wildlands	Hearthstone
7	Bejeweled	Civilization	For Honor	Heroes of the Storm
8	Battlefront	The Darkness	Tom Clancy's	
9	NBA	XCOM	The Division	
10	Dragon Age	WWE		
11	Titanfall	GTA		
12	Dead Space	Grand Theft Auto		
13		Max Payne		
14		Red Dead Redemption		
15		RRD		

1.3 Technical Analysis

1.3.1 Moving Averages

Several types of moving averages are used. The first and simplest of them is a simple moving average (SMA). It takes into account in the same degree all observations in a given time window. For obvious reasons, one may assume that closer dates may have more significant influence on future price. Therefore, in addition to the SMA, Weighted

Moving Average (WMA) and Exponential Moving Average were used. The WMA solves the aforementioned problem by giving more weight to more recent data. EMA works in a similar way, but the price change is not consistent but exponential.

1.3.2 Bollinger Bands

Bollinger Bands consist of three bands. The middle one is a moving average. Higher and lower bands are deviated from the middle one by 2 standard deviations up and down respectively.

1.3.3 Bollinger Bands

Moving Average Convergence Divergence (MACD) consists of two lines. The core of the indicator is the MACD line which is the difference between the 12-period EMA and the 26-period EMA. The second line is the signal line which is a 9-period EMA. Their position relative to each other helps to determine whether the market is oversold or overbought.

1.4 Data Scaling

Due to the use of neural network based architectures in the study, rescaling of the data was required. For this purpose, min-max normalization was used. The equation for the rescaled value is:

$$x_{\text{scaled}} = \frac{x - min(x)}{max(x) - min(x)}$$

1.5 Sentiment Analysis

Sentiment analysis is a process of extracting users' feelings and emotions. It is a part of Natural Language Processing. It boils down to trying to determine with a given probability whether a given statement was positive, negative or neutral. Then such predictions are converted into numerical data. There are very many different types of models for sentiment analysis, but the vast majority of them are based on machine learning. In our study, the BERT (Bi-Directional Encoder Representations from Transformers) model created by Google researchers was used.

1.6 BERT (Bi-Directional Encoder Representations from Transformers)

BERT is a state-of-the-art NLP model. One of its biggest advantages is taking whole sentences as an input in contrast to traditional NLP models that take one word a time. One of BERT's biggest advantages is that it is a semi-supervised model. It is pre-trained on very large sets of non labeled data, learning to fill gaps in the text. Then this model can be trained for any task just by adding one extra layer. Moreover, one can find many pre-trained models ready for download on the internet. This makes it possible for ordinary users without astronomical computational capabilities to use such a powerful model for their tasks by training it only on small amounts of labeled data.

2 Models

2.1 Basic Recurrent Neural Network

Recurrent neural networks (RNN) is an extended version of artificial neural networks. Its main advantage is having internal memory which allows it to process sequences. Hidden state allows previous outputs to be used as input for further parts of the sequence. This ability to remember previous states makes RNN suitable for time series forecasting.

2.2 Gated Recurrent Unit

Gated Recurrent Unit is an extended version of Recurrent Neural Networks which adress RNN vanishing gradient problem [2]. It's architecture has been proposed by K.Cho in 2014 [3].

References

[1] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, 2014

[2] P. Dey et al., Comparative Analysis of Recurrent Neural Networks in Stock Price Prediction for Different Frequency Domains, 2021

[3] K. Cho et al., On the Properties of Neural Machine Translation: Encoder-Decoder Approaches, 2014