isrjournals

International Journal of Advanced Research in Computer Science Engineering and Information Technology

Volume: 6, Issue: 3,Special Issue: 1 ,Apr,2021 ,ISSN_NO: 2321-3337

RECOMMENDATION SYSTEM USING DEEP LEARNING

G. Anbuselvi M. Tech(Ph.D), Professor, Department of M.E (Computer Science and Engineering),

A.P. Preetha, Student, Department of M.E (Computer science and engineering),

Meenakshi College of Engineering, India.

Abstract

Investment firms, hedge funds and even individuals have been using financial models to better understand market behavior and make profitable investments and trades. A wealth of information is available in the form of historical stock prices and company performance data, suitable for machine learning algorithms to process.Can we actually predict stock prices with machine learning? Investors make educated guesses by analyzing data. They'll read the news, study the company history, industry trends and other lots of data points that go into making a prediction. The prevailing theories is that stock prices are totally random and unpredictable but that raises the question why top firms like Morgan Stanley and Citigroup hire quantitative analysts to build predictive models. We have this idea of a trading floor being filled with adrenaline infuse men with loose ties running around yelling something into a phone but these days they're more likely to see rows of machine learning experts quietly sitting in front of computer screens. In fact about 70% of all orders on Wall Street are now placed by software, we're now living in the age of the algorithm.

Keywords—prediction; forecast, datamining , recommendation

I. INTRODUCTION

With the explosive growth of social media (e.g. blogs, micro-blogs, forum discussions and reviews) in the last decade, the web has drastically changed to the extent that nowadays billions of people all around the globe are freely allowed to conduct many activities such as interacting, sharing, posting and manipulating contents. This enables us to be connected and interact with each other anytime without geographical boundaries, as opposed to the traditional structured data available in databases. The resulted unstructured user-generated data mandates new computational techniques from social media mining, while it provides us opportunities to study and understand individuals at unprecedented scales [1, 2, 3, 4, 5, 6, 7]. Sentiment analysis (a.k.a opinion mining) is one class of computational techniques which automatically extracts and summarizes the opinions of such immense volume of data which the average human reader is unable to process. This ocean of opinionated postings in social media is central to the individuals' activities as they impact our behaviors and help reshape businesses. Nowadays, not only individuals are no longer limited to asking friends and family about products but also businesses, organizations and companies do not require to conduct surveys or polls for opinions about products, as there are tons of user reviews and discussions in public forums on the

Web. There are thus numerous immediate and practical applications and industrial interests of collecting and studying such opinions by using computational sentiment analysis techniques, spreading from consumer products, services, healthcare, and financial services to social events, political elections and more recently crisis management and natural disasters.

Social media has pervasively played an increasing role and they have become an important alternative information channel to traditional media in the last five years during emergencies and disasters, where they rank as the fourth most popular sources to access necessary information during emergencies [8, 9]. In particular, individuals and communities have used social media for many tasks from warning others of unsafe areas to fund raising for disaster relief [8]. The days of one-way communication where only official sources used to provide bulletins during emergencies are actually gone. In 2005 for instance, when Hurricane Katrina slammed the U.S. gulf coast, there was no Twitter for news update while Facebook was not that much famous yet. Compare, for example, Hurricane Katrina to the Haiti earthquake on January 2010. During latter, people used Twitter, Facebook, Flicker, blogs and YouTube to post their experience in form of texts, photos and videos during the earthquake which resulted in donating 8 million U.S. dollars to the Red Cross which vividly demonstrates the power of social media in propagating information during emergencies [10].Hurricane Sandy on 2012, is another example to show the positive impact of social media during disasters. By that time, using social media had become an important part of disaster response. There are numerous similar examples that show how social media have come to the rescue in disaster situations including Hurricane Irene, California gas explosion on 2010, Japan earthquakes, Genoa flooding and more recently Ebola. Social media could be actually leveraged to keep the problem informed, help locate loved ones, and express support or notify authorities during emergencies and disasters. Sentiment analysis of disaster related posts in social media in could help to detect posts that contribute to the situational awareness and better understand the dynamics of the network including users' feelings, panics and concerns by identifying the polarity of sentiments expressed by users during disaster events to improve decision making. Sentiment information could also be used to project the information regarding the devastation and recovery situation and donation requests to the crowd in better ways.

Interactive tools such as visual analytic methods could help us to make a large amount of complex information more readable and interpretable, if integrated by computational approaches, as the effectiveness of most computational techniques is limited due to several factors [11]. Interactive visual analytics provide intuitive ways of making sense of large amount of posts available in social media. These techniques are now widely used in social media data and contribute in many areas of exploratory data analysis. Despite most social media visualization approaches which rely solely on geographical and temporal features, there are some systems which are able to exploit the sentiments of the data such which help improving visualization. Besides disaster related data management in social media, the ability to drawing out important features could be used for better and quick understanding of situation which leads to rapid decision making in critical situations. Moreover, the data produced by social media

during disasters and events, is staggering and hard for an individual to process. Therefore, visualization is needed for facilitating pattern discovery.

II. OVERVIEW

Investment firms, hedge funds and even individuals have been using financial models to better understand market behavior and make profitable investments and trades. A wealth of information is available in the form of historical stock prices and company performance data, suitable for machine learning algorithms to process.

Can we actually predict stock prices with machine learning? Investors make educated guesses by analyzing data. They'll read the news, study the company history, industry trends and other lots of data points that go into making a prediction. The prevailing theories is that stock prices are totally random and unpredictable but that raises the question why top firms like Morgan Stanley and Citigroup hire quantitative analysts to build predictive models. We have this idea of a trading floor being filled with adrenaline infuse men with loose ties running around yelling something into a phone but these days they're more likely to see rows of machine learning experts quietly sitting in front of computer screens.In fact about 70% of all orders on Wall Street are now placed by software, we're now living in the age of the algorithm. This project seeks to utilize Deep Learning models, Long-Short Term Memory (LSTM) Neural Network algorithm, to predict stock prices. For data with timeframes recurrent neural networks (RNNs) come in handy but recent researches have shown that LSTM, networks are the most popular and useful variants of RNNs.

III. LITERATURE REVIEW

Literature survey is a review of an abstract accomplishment. It provides a way to study and get a clear understanding of critical points of current knowledge including substantive findings as well as theoretical and methodological contributions to a particular topic.

Amnon Shashua (2005)

Presented an algebraic approach to variable weighting, which is based on maximizing a score based on the spectral properties of the kernel matrix. The approach has the advantage of being suitable to unsupervised feature selection, but can also be applied in the supervised settings. It is interesting to compare the algebraic approach presented in this work to probabilistic approaches which take a ”holistic” view of the data such as the information bottleneck and the info max. The gap that exists between the probabilistic tools of machine learning makes a direct comparison to information-based Feature selection criteria a subject for future work.

George Forman(2003)

The residual analysis determined that BNS paired with Odds Ratio yielded the best chances of attaining the best precision. For optimizing recall, BNS paired with F1 was consistently the best pair by a wide margin. Future work could include extending the results for nominal and real-valued feature values, and demonstrating BNS for non-text domains. The feature scoring methods we considered are oblivious to the correlation between features; if there were ten duplicates of a predictive feature, each copy would be selected.

Zheng Zhao(2007)

This work presents a concrete initial attempt to the new problem of semi- supervised

feature selection. We propose an algorithm based on the spectral graph theory.

Feiping Nie, Xiao Cai, (2004)

Motivated by previous work, the `2;1-norm regularization is used to select features across all data points with joint sparsity. We provided an efficient algorithm with proved convergence. Our method has been applied into both genomic and proteomic biomarkers discovery.

Deng Cai, Chiyuan Zhang, and Xiaofei He(2010)

In comparison with one simple method, that is, Max Variance, and two state-of-the-art methods, namely, Lap lamina Score , the experimental results validate that the new method achieves significantly higher performance for clustering and classification. Our proposed MCFS algorithm performs especially well when the number of selected features is less than 50.

Michael J. Lyons, Julien Budynek, and Shigeru Akamatsu, december (1999)

The saliency information displayed in shows that theregions around the eyes and mouth are more important than other areas of the face for classifying the facial expressions. Filters of intermediate spatial frequency were found to be slightly more informative for expression classification. Notably, filters having horizontal orientation were more heavily weighted in the discriminant vector than other orientations. This seems intuitively correct since the most noticeable expressive motions of the face are the opening and closing of the mouth and eyes and raising and lowering of the eyebrows. Displacement of roughly horizontal edges forms the largest component of these motions.

Lior Wolf and , Amnon Shashua(2004)

In this work we presented an algebraic approach to variable weighting, which is based on maximizing a score based on the spectral properties of the kernel matrix. The approach has the advantage of being suitable to unsupervised feature selection, but can also be applied in the supervised settings. It is interesting to compare the algebraic approach presented in this work to probabilistic approaches which take a ”holistic” view of the data such as the information bottleneck and the infomax.The gap that exists between the algebraic and the probabilistic tools of machine learning make a direct comparison to information-based feature selection criteria a subject for future work.

Zhigang Ma, Feiping Nie, Yi Yang and Jasper Uijlings(2010)

To validate the efficacy of our method for web image annotation, we conducted experiments on two popular image databases consisting of web images. It can be seen from the experimental results that our method outperforms classical and state of the art algorithms for image annotation. Therefore, we conclude that our method is a robust feature selection method and its feature subspace sharing foundation makes it particularly suitable for web images which are usually multilabeled.

Zechao and Jing Liu(2003)

We proposed the MPMF model for image annotation, which integrates the image-word correlation, image similar- ity and word correlation simultaneously and seamlessly. Different from standard models, MPMF connects these three di_erent data resources through the shared word latent feature space and the shared image latent feature space. The experiments on the Corel dataset and the Flicker image dataset emonstrate that the

proposed approach is more preferable than the state-of-the-art algorithms.Spam detection is still extensively investigated in Web-Web and E-mail domains (Gyo¨ngyi et al., 2004; Ntoulas et al., 2006), while research has recently been expanded into the domain of customer reviews [2](Different types of indicator signals have been investigated.For example, trained Jindal and Liu (2008)models use content-based features to review, review, and the product itself. Yoo and Gretzel (2009) compiled a review of 40 authentic and 42 fake hotels and manually compared the language differences between them.

Ott et al. (2011)

Created a database of ratings by recruiting Turkers to write false reviews. Their data are accepted by the line of work that follows (Ott et al., 2012; Feng et al., 2012; Feng and Hirst, 2013). For example, Feng et al. (2012) looked at syntactic materials from Context Free Grammar (CFG) cleaning trees to improve performance. Feng and Hirst (2013) create hotel profiles from clusters of reviews, measures the relevance of customer reviews on a hotel profile, and uses it as a feature of spam detection. Recently, Li et al. (2014) created a broad integration benchmark, which included data from three domains (Hotel, Restaurant, and Surgeons), and explored common ways of identifying spam for viewing ideas online. We accept this data for our experiments because of its large size and integration.

Existing methods use traditional syntactic elements, which can be small and fail to incorporate semantic information from complete speech. In this paper, we propose to study the representation of the neural levels of a document to better identify spam ideas. To the best of our knowledge, we are the first to investigate the intensive education of spam detection of delusional ideas.There is some work to do without the content of the review itself.

In addition to Jindal and Liu (2008), Mukherjee et al. (2013) examined factors from customer behavior to detect fraud. Based on factual reviews and numerous unlisted reviews, Ren et al. (2014) proposed a supervised learning approach, and created an intuitive classifier to detect deceptive updates. Kim et al. (2015) introduced an independent semantic-based feature based on FrameNet. Experimental results indicate that semantic independent features can improve classification accuracy.

Neural network models have been misused to study the dense feature representation for a variety of NLP functions (Collobert et al., 2011; Kalchbrenner et al., 2014; Ren et al., 2016b). Distributed word returns (Mikolov et al., 2013) have been used as a basic building block with many NLP models. Numerous methods have been proposed to study the introductions of phrases and large sections of texts from the vocabulary distribution. For example, Le and Mikolov (2014) introduced a vector of categories to read document presentations, extending the word embedding methods of Mikolov et al. (2013). Socher et al. (2013) introduced a family of recuralive neural recommivation networks to represent a semantic level category. Subsequent research includes a multidimensional network of neural and global feed back.

IV PROPOSED SYSTEM

The challenge of this project is to accurately predict the future closing value of a given stock across a given period of time in the future.

For this project I will use a **Long Short Term Memory networks**[1] – usually just called "LSTMs" to predict the closing price of the **S&P 500**[2] using a dataset of past prices.

GOALS

1. Explore stock prices.
2. Implement basic model using linear regression.
3. Implement LSTM using keras library.
4. Compare the results and submit the report.

METRICS

For this project measure of performance will be using the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) calculated as the difference between predicted and actual values of the target stock at adjusted close price and the delta between the performance of the benchmark model (Linear Regression) and our primary model (Deep Learning).

Data Exploration

The data used in this project is of the **Alphabet Inc**[3] from **January 1, 2005 to June 20, 2017**,this is a series of data points indexed in time order or a time series. My goal was to predict the closing price for any given date after training. For ease of reproducibility and reusability, all data was pulled from the **Google Finance Python API**[4].The prediction has to be made for Closing (Adjusted closing) price of the data. Since Google Finance already **adjusts the closing prices for us**[5], we just need to make prediction for "CLOSE" price.

The dataset is of following form:

Date	Open	High	Low	Close	Volume
30-Jun-17	943.99	945.00	929.61	929.68	2287662
29-Jun-17	951.35	951.66	929.60	937.82	3206674
28-Jun-17	950.66	963.24	936.16	961.01	2745568

Table: The whole data can be found out in 'Google.csv' in the project root folder

Note: I did not observe any abnormality in datasets, i.e, no feature is empty and does not contains any incorrect value as negative values.

Feature	Open	High	Low	Close	Volume
Mean	382.5141	385.8720	378.7371	382.3502	4205707.8896
Std	213.4865	214.6022	212.08010	213.4359	3877483.0077
Max	1005.49	1008.61	1008.61	1004.28	41182889
Min	87.74	89.29	86.37	87.58	521141

We can infer from this dataset that date, high and low values are not important features of the data. As it does not matter at what was the highest prices of the stock for a particular day or what was the lowest trading prices. What matters is the opening price of the stock and closing prices of the stock. If at the end of the day we have higher closing prices than the opening prices that we have some profit otherwise we saw losses. Also volume of

share is important as a rising market should see rising volume, i.e, increasing price and decreasing volume show lack of interest, and this is a warning of a potential reversal. A price drop (or rise) on large volume is a stronger signal that something in the stock has fundamentally changed.Therefore, I have removed Date, High and low features from data set at preprocessing step. The mean, standard deviation, maximum and minimum of the preprocessed data was found to be following:

	Mean	Std	Max	Min
Open	0.3212	0.23261	1.0	0.0
Close	0.3215	0.2328	1.0	0.0
Volume	0.09061	0.0953	1.0	0.0

Through this data we can see a continuous growth in Alphabet Inc. The major fall in the prices between 600-1000 might be because of the Global Financial Crisis of 2008-2009.

Algorithms and Techniques

The goal of this project was to study time-series data and explore as many options as possible to accurately predict the Stock Price. Through my research i came to know about **Recurrent Neural Nets (RNN)[8]** which are used specifically for sequence and pattern learning. As they are networks with loops in them, allowing information to persist and thus ability to memorise the data accurately. But Recurrent Neural Nets have vanishing Gradient descent problem which does not allow it to learn from past data as was expected.

The remedy of this problem was solved in **Long-Short Term Memory Networks[9]**, usually referred as LSTMs. These are a special kind of RNN, capable of learning long-term dependencies.

In addition to adjusting the architecture of the Neural Network, the following full set of parameters can be tuned to optimize the prediction model:

Input Parameters

- Preprocessing and Normalization (see Data Preprocessing Section)

Neural Network Architecture

- Number of Layers (how many layers of nodes in the model; used 3)
- Number of Nodes (how many nodes per layer; tested 1,3,8, 16, 32, 64, 100,128)

Training Parameters

- Training / Test Split (how much of dataset to train versus test model on; kept constant at 82.95% and 17.05% for benchmarks and lstm model)
- Validation Sets (kept constant at 0.05% of training sets)
- Batch Size (how many time steps to include during a single training step; kept at1 for basic lstm model and at 512 for improved lstm model)
- Optimizer Function (which function to optimize by minimizing error; used “Adam” throughout)
- Epochs (how many times to run through the training process; kept at 1 for base model and at 20 for improved LSTM)

isrjournals

International Journal of Advanced Research in Computer Science Engineering and Information Technology

Volume: 6, Issue: 3,Special Issue: 1 ,Apr,2021 ,ISSN_NO: 2321-3337

VI RESULTS AND DISCUSSION

With each model i have refined and fined tune my predictions and have reduced mean squared error significantly.

For my first model using linear regression model:

- **Train Score: 0.1852 MSE (0.4303 RMSE)**
- **Test Score: 0.08133781 MSE (0.28519784 RMSE)**

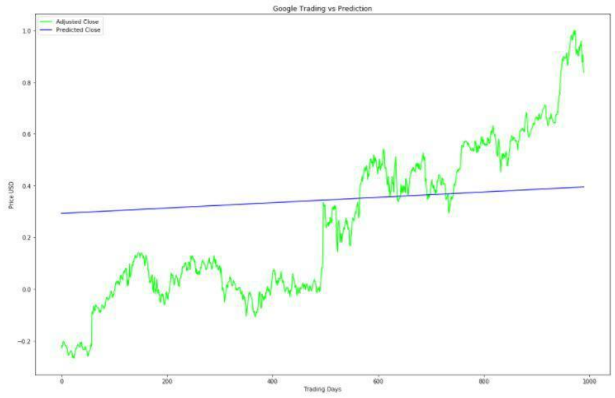


Fig: Plot of Linear Regression Model

For my second model using basic Long-Short Term memory model:

- **Train Score: 0.00089497 MSE (0.02991610 RMSE)**
- **Test Score: 0.01153170 MSE (0.10738577 RMSE)**

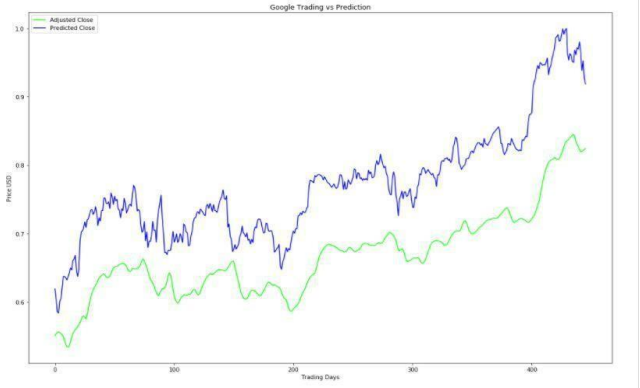


Fig: Plot of basic Long-Short Term Memory model

For my third and final model, using improved Long-Short Term memory model:

- **Train Score: 0.00032478 MSE (0.01802172 RMSE)**
- **Test Score: 0.00093063 MSE (0.03050625 RMSE)**

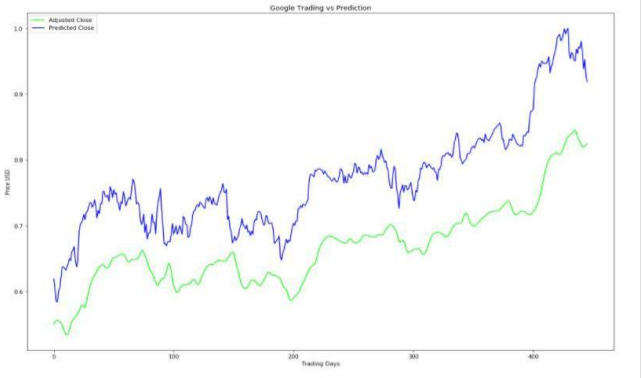
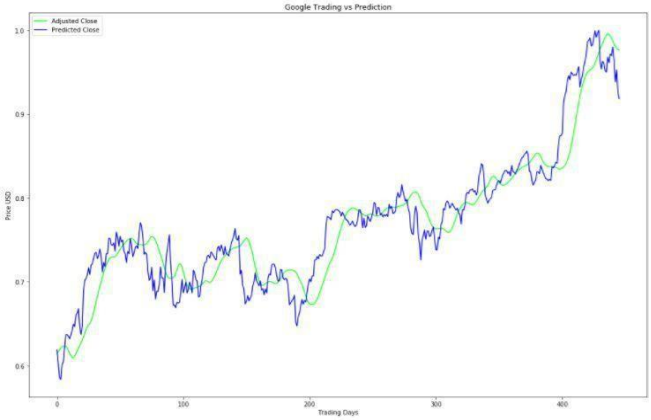


Fig: Plot of Improved Long-Short Term Memory Model

Robustness Check:

For checking the robustness of my final model i used an unseen data, i.e, data of Alphabet Inc. from July 1, 2017 to July 20, 2017. On predicting the values of unseen data i got a decent result for the data. The results are as follows:



Test Score: 0.3897 MSE (0.6242 RMSE)

Comparing the benchmark model - Linear Regression to the final improved LSTM model, the

Mean Squared Error improvement ranges from **0 .08133781 MSE (0.28519784 RMSE)** [Linear Regression Model]to**0.00093063 MSE (0.03050625 RMSE)**[Improved LSTM].This significant decrease in error rate clearly shows that my final model have surpassed the basic and benchmark model.Also the Average Delta Price between actual and predicted Adjusted Closing Price values was:

Delta Price: 0.000931 - RMSE * Adjusted Close Range

VII CONCLUSION

I started this project with the hope to learn a completely new algorithm, i.e, Long-Short Term Memory and also to explore a real time series data sets. The final model really exceeded my expectation and have worked remarkably well. I am greatly satisfied with these results.

The major problem I have faced during the implementation of project was exploring the data. It was toughest task. To convert data from raw format to preprocess data and then to split them into training and test data. All of these steps require a great deal of patience and very precise approach. Also I had to work around a lot to successfully use the data for 2 models, i.e, Linear Regression and Long-Short Term Memory, as both of them have different inputs sizes. I read many research papers to get this final model right and i think it was all worth it :)

VII FUTURE WORK

The biggest challenge for e-commerce businesses is ensuring a superior customer service to shoppers. Helping them find what they are looking for and guiding their shopping experience is what makes the process challengeable. In brick-and-mortar stores, you can always find savvy salespeople. They help to find what the shopper looks for and gives specific recommendations based on their preferences and wishes.

VIII REFERENCE

[1] T. Takahashi and N. Igata. Rumor detection on Twitter. In 6th International Joint Conference SCIS and ISIS pages 452- IEEE ,2012

[2] J. Ronson. So you've Been Publicly shamed. Picador, 2015

[3] Y R Tausczik and J W Pennebaker. The psychological meaning of words: LIWC and computerized text analysis method. Journal of language and social Psychology, 29(1):24-54, 2010

[4] C-C. Chang and C.J Lin, "LIBSVM: A library for support vector machines." ACM Trans . Intell. syst. Technol ., vol. 2, no.3, pp. 27:1-27:27, 2011

[5] Hong and S.H. Kim, "Political polarization on Twitter: Implications for the use of social media in digital government." 2016

[6] Blockshame shields you from the online Mob Just in case! Accessed : Feb 7,2018

[7] Twitter Report Abusing Behavior. Accessed: Feb 7,2018

[8] S.Rojas - Galeano, " On obstructing obscenity obfuscation,"ACM Trans. Web, vol. 11,no.2, p. 12, 2017

[9] Hate-Speech. Oxford Dictionaries. Accessed: Aug. 30, 2017

[10] I. Kwork and Y. Wang, "locate the hate: Detection tweets against blacks" in Proc. AAAI, 2013