Math 7243-Machine Learning-Final Project: Predicting VIX Ticker Volatility from Headlines

Josh Galloway (galloway.j@husky.neu.edu),
Northeastern University

April 2020

Abstract

Working from publicly available data sources, the student trained a recursive neural network (RNN) to predict the Chicago Board Options Exchange (CBOE) Volatility Index (VIX). An ad hoc method of sentiment quantification was applied to data sets containing Reddit and ABC headlines spanning several years. The result was a trained RNN that was able to predict the Δ VIX closing day over day based on the headlines from the previous 20 trading days.

1 Introduction

The utilization of machine learning techniques to predict stock market metrics based on headlines is not novel; however, several challenges still persist. Here, the target metric to predict was the VIX. The VIX is an indication of the expected fluctuation in the price (volatility) of the SP 500's index options over a forward-looking 30 day period [Kue]. The dataset was compiled and formatted several different ways until arriving upon the most performant structure. Preprocessing of both the inputs and output/target data for the model was necessary; in addition, the structure and hyperparameter spaces for the RNN were programmatically explored for selection of the most effect solutions.

2 Dataset

Inputs to the model were derived from the news headline data sets publicly available. The sets used here were for ABC news [Roh] and Reddit WorldNews Channel [Aar]. The ABC news dataset is comprised of around 250 headlines per day ranging from dates 19 Feb. 2003 to 31 Dec. 2019. The Reddit dataset is comprised of the top 25 headlines for each day from 8 June 2008 to 01 July 2016.

The VIX data was obtained from the CBOE website and the data archive utilizing the more modern method of calculating the metric was utilized [VIX]. This data contained the daily VIX value's high and low, as well as open and closing. The sources were readily available and already partially cleaned for use in a machine learning project. A more exhaustive and varied dataset would be desirable to improve the robustness of the model; however, it is a highly labour intensive process to gather such a dataset. A summary of the datasets and test/train split can be found in table 1.

Dataset (Samples)	**Description**	**Train/Test Split**	**Input Sample Shape**
Full Dataset (1989)	Dates where both ABC and Reddit headlines were available.	80%/20%	(20,3)
ABC (4238)	All ABC headlines		
Reddit (1989)	All Reddit headlines		
Combined (6227)	All Reddit headlines appended to all ABC headlines		

Table 1: Dataset Size and Splits

Conditioning the headline data was an exercise in sentiment analysis. To this end, the raw headline text was scored for sentiment by means of an ad hoc method. This involved processing each headline and referencing against selected portions of the Harvard IV-4 dictionary and Lasswell value dictionaries[SEN]. This dictionary provides scores for each word in binary classification for complimentary terms such as Positive/Negative, Pleasure/Pain and so on. As mentioned, only three of these pairs were utilized as they were the base pairs off of which several subcategories were derived. The utilized categories were Positive/Negative, Strong/Weak, Active/Passive.

3 Data Representation and Processing

For the headline data, the values were condensed to produce a three-dimensional sentiment score. The first dimension was chosen to be Positive/Negative sentiment taking on binary values of 1/-1, respectively. Similarly, Strong/Weak and Active/Passive were mapped to the other two dimensions producing a final vector for the sentiment of each headline [fig. 1]. This resulting vector was then converted to a unit vector in order to not penalize shorter headlines or favor headlines that contained more of the words found in the dictionary than others by chance. The sentiment score was calculated in several manners and cycled through to empirically determine the closest to optimal scoring method and training set. A cumulative score, average score and unit vector of each day's score. This was done for the source datasets separately, (an ABC set of scores and Reddit Set of Score) and then for the sum of both daily scores creating a combined score. The VIX target data in raw form had a tail heavy distribution. In order to correct this and make the data more amenable to machine learning, the Δ each daily metric was taken. The result more closely resembled a Gaussian-shaped distribution [fig. 2].

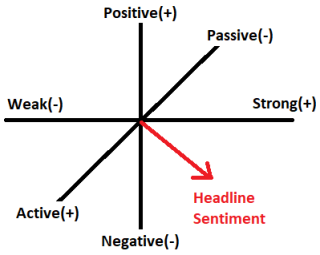

Figure 1: Sentiment Vector

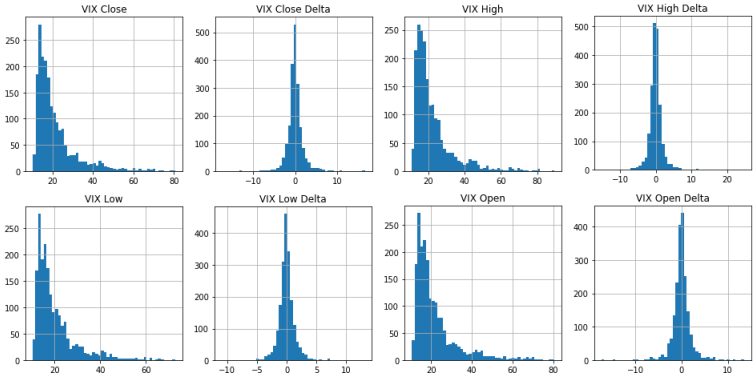

Figure 2: VIX Data Conditioning

4 Model Derivation

Starting with the cleaned and conditioned dataset, several steps followed to build a working model. The dataset was visualized to try to gain any insights as to which metrics would work best. An RNN was selected to model the time series data, as this is its main field of application. Test models of shallow training depth were built in further exploration and the field narrowed successively to produce a final metric and feature set to model. Then, methods of presenting the datasets to the RNN were empirically examined to arrive at a training methodology. Finally, the structure and hyperparameter space were explored via grid searches and the ultimate model was trained with the optimized parameters.

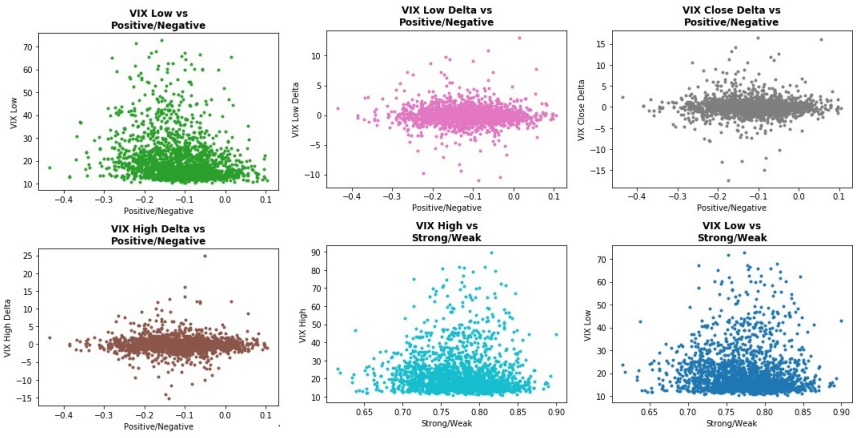

Figure 3: Initial Dataset Visualization

Beginning to create the model started with visualizing the dataset. All of the data set dimensions were graphed against the conditioned and unconditioned VIX data. The visualization points to the use of the conditioned (i.e. Δ VIX) metric over the raw data as its increased correlation is readily apparent over the unconditioned metrics [fig. 3]. Exploratory models of the form in table [tab. 2] for all 72 model combinations. These were 8 targets (VIX Close, Open, Daily High, Daily Low and the Δ of each) trained from 9 features–summation, average and unit sentiment vectors for each of the 3 headlines sets (ABC, Reddit and then the set where both headlines were combined daily). The models were trained for 50 epochs with a mini-batch size of 32 and evaluated against a loss function of mean-squared error with a validation split of 10%. A step size for slicing the time series data of 20 trading days (one month) was chosen as it is the basis for the VIX metric.

LSTM	**Dropout Layer**	**Dense Layer**	**Batch Normalization**	**Activation Layer**	**Optimizer**
Nodes = 64, Input = (20,3)	20%	Nodes = 1, ELU, HE Normal	–	ELU, HE Normal	Adam, LR = 0.001

Table 2: Exploratory Model Structure

The best performing models from this round of exploration were for the Δ VIX closing metric with the ABC dataset out performing the Reddit and cumulative (ABC and Reddit summed per day) sets. However, the shape of the learning curves suggested the ABC data was leveling off (velocity becoming less negative) and the slope of the other learning curves presented a more linear slope (constant negative velocity). This led to further training for 250 epochs of the remaining 9 models to decide on the final model features. The Δ VIX closing metric was solidified at this point as the target variable. The result of deeper training is shown in [fig. 4], and leads to the conclusion that training on the datasets for Reddit and ABC separately is preferable to the additive feature set engineered from their daily synthesis.

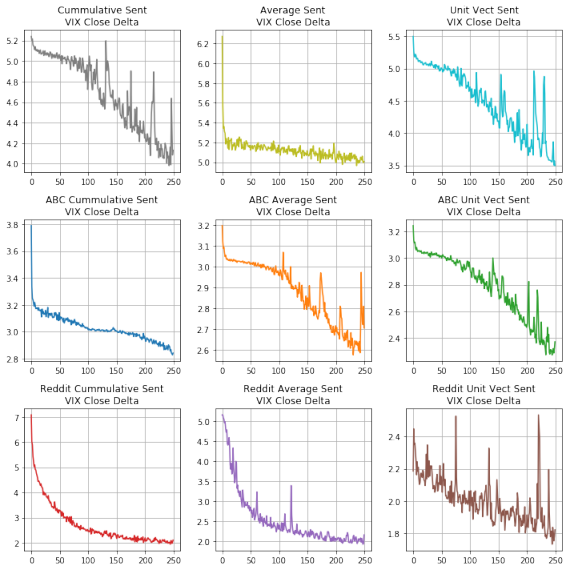

Figure 4: Best Nine Models Training for Loss vs 250 Epochs

This led to the decision to train the network on new "Combined" dataset which simply appended the Reddit dataset to the end of the ABC dataset. Again, this new dataset was trained for 250 epochs resulting in [fig. 5]. Cumulative sentiment for the combined dataset with Δ VIX at closing as the target was the best performing model. An argument could be made that the slope of the unit vector feature's learning curve is steeper and could be worth further exploration, but the instability of the curve lead to the final selection of the previously mentioned pairing.

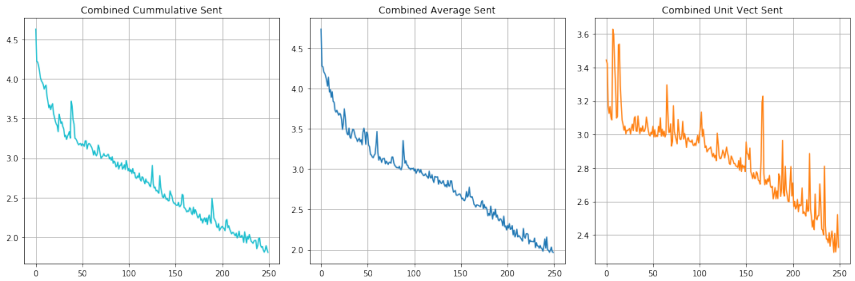

Figure 5: Loss for VIX Δ Closing Trained on Reddit Dataset Appended to ABC

The cell-type, structure, and hyper-parameters of the network were determined via three total grid searches with shallow training sessions. The search criteria and results are in [tab. 3].

Grid Search	**Parameters Searched**	**Parameters Held Constant**	**Resulting Best**
#1	Node Type = [LSTM, GRU] No. Nodes =[32, 64, 128, 256, 512] Dropout = [0.1, 0.2, 0.4]	Epochs = 10, Batch Size = 23, Learning Rate = 0.001	128 LSTM Nodes, Dropout = 0.2
#2	Epochs = [10,100,250,500] Batch Size = [16,32,64,1024, Full Batch (4943)]	128 LSTM Nodes, Dropout = 0.2, Learning Rate = 0.001	Epochs = 250, Full Batch
#3	Learning Rate = [0.0001, 0.0012, 0.0023, 0.0034, 0.0045, 0.0056, 0.0067, 0.0078, 0.0089, 0.01]	128 LSTM Nodes, Dropout = 0.2, Epochs = 250, Full Batch	Learning Rate = 0.0067

Table 3: Hyperparameter Optimization

The final model was selected from the optimized structure and hyperparameters and trained with the combined dataset features with a target of Δ VIX closing for the regression. The final training epochs were optimized by hand starting with 1000 epochs and working toward an equal value of validation-loss and training-loss. This phenomenon should roughly signify the point at which the RNN begins overfitting to the training set, and it occurred at about 310 epochs [fig. 6]

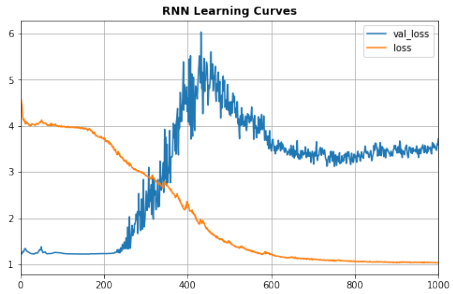

Figure 6: Final Model Learning Curves

5 Results

Once trained, the network was scored against a test set to root-mean-squared error (RSME), and evaluated again by graphing the results of the prediction against the recorded data set. The RMSE was 1.87 for the ABC dataset against the segregated test set, and 2.07 for the Reddit. The predictions graphed against both the ABC dataset [fig. 7] and the Reddit dataset [fig. 8] are shown.

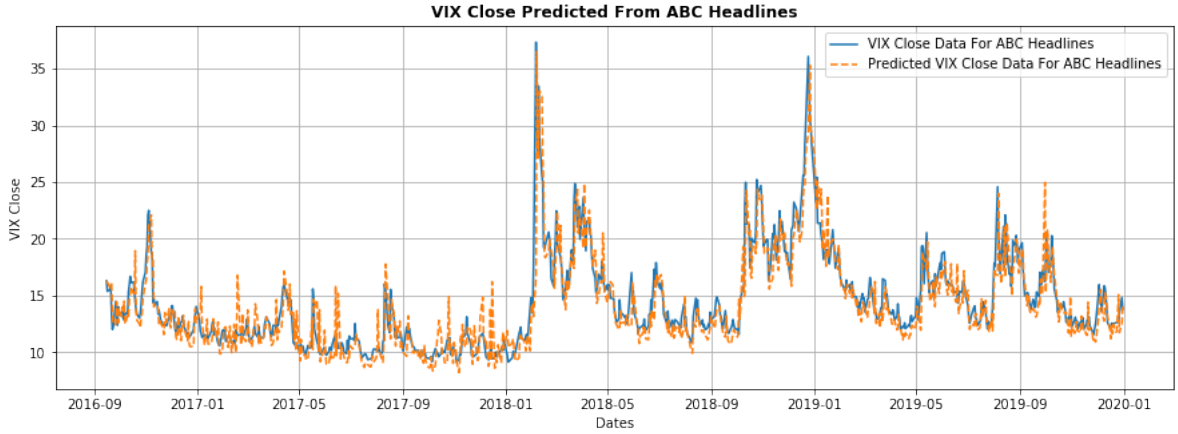

Figure 7: Cumulative Headline Score for ABC Dataset Modeling Δ VIX Closing

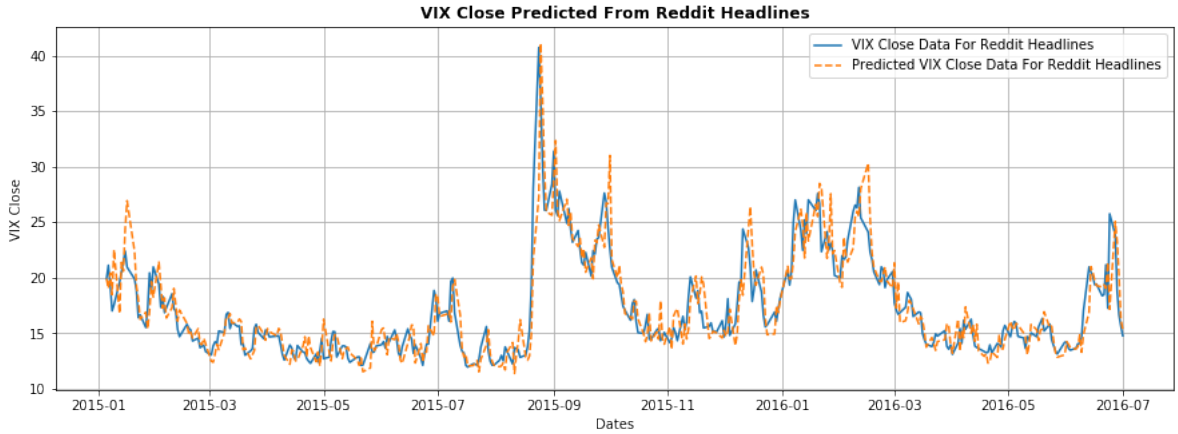

Figure 8: Cumulative Headline Score for Reddit Dataset Modeling Δ VIX Closing

6 Conclusion

Overall, the model performed well as evinced by the graphing against the target data. There were, however, some inaccuracies during periods of low volatility in which the model over-predicted the target value producing some erroneous spikes. However, an overestimation may be preferred in some cases such as a more risk averse trading strategy. Further improvements could likely be made by diversification of the new sources and general expansion of the training set. Other methods of sentiment analysis could also be explored utilizing other machine learning techniques such as support vector machines and random forest classifiers.

References

[Aar] Aaron7sun. Daily news for stock market prediction. Retrieved from https://www.kaggle.com/aaron7sun/stocknews/data.

[Kue] J. Kuepper. Cboe volatility index (vix) definition. Retrieved from https://www.investopedia.com/terms/v/vix.asp.

[Roh] Rohk. A million news headlines. Retrieved from https://www.kaggle.com/therohk/million-headlines.

[SEN] Welcome to the general inquirer home page. (n.d.). Retrieved from http://www.wjh.harvard.edu/ inquirer/.

[VIX] Vix index historical data. (n.d.). Retrieved from http://www.cboe.com/products/vix-index-volatility/vix-options-and-futures/vix-index/vix-historical-data.