HIGH SCHOOL EDITION
Journal of Student Research
Volume 12 Issue 3 (2023)

News Never Knows

Krish Patil[1] and Philip Bell[2#]

[1]Del Norte High School
[2]Inspirit AI
[#]Advisor

ABSTRACT

Finding a possible correlation between the news and how the stock market reacts to it is important as it would allow people to predict how the stock market will change in the future, allowing investors to make smarter decisions with their money and make sure that they are getting the maximum profit possible. People at risk could benefit from a prediction system because they could know how the stock market is going to change before it actually happens, causing them to avoid something bad happening. Along with this, businesses looking to other businesses for possible partnerships will be able to determine if having a partnership with a certain business will be profitable or not. A prediction system would allow us to overcome people losing massive amounts of money as they would have a prediction system stopping them from making unprofitable decisions in the stock market.

INTRODUCTION

The goals of this research is to determine whether a relationship exists between news headlines and how the stock market performs. If a relationship like this does exist, we would like to identify what exactly this relationship is. Along with this, a goal is to determine what the best model to predict whether the stock market will go up or down is in this scenario. The reason this research is being conducted is that it is important to see how the stock market is affected by the news that's put out daily. Figuring out how the stock market reacts to news can help us figure out how the economy works. People at risk can benefit from a prediction system as they would know how the stock market is going to change before it actually happens, causing them to avoid possible disaster and saving their money from being lost. Some barriers that could be overcome is making people aware of the risks that they're taking when they make a certain investment.

LITERATURE REVIEW

When going into the topic at first, an artice was found authored by Lee and Timmons at Stanford University in which they tested the relationship between news articles and how the stock market reacts [1]. They developed an AI model that predicts whether a word is positive or negative within the article and based on that, told what to do in the stock market. While the baseline method only gave them a 0.615 % per month return, their new method gave them a 2.05% return per month. Although the best results are higher, the comparison is inconsistent. This suggested that there isn't a high correlation between the news and the stock market, though we wanted to try test that theory for ourselves. Since people are more likely to simply skim over a title rather than read the entire article in depth, the research conducted was focused on the titles of news articles and how the stock market reacts.

METHODS

To discover if there is a relationship between how the stock market performs and the top news headlines from that day, we used a pre-trained model within a pipeline which took a list of words and gave them a sentiment between -1 and 1, -1 being

most negative and 1 being most positive. Then, we added up all of the sentiments of the top 25 articles from that day and reported that as the final overall sentiment value for that day. We then used that overall sentiment value and the provided data of whether the stock market went up or down on that day to train our AI. The data was split up into training and testing sets so that the models could be tested to see how accurate they are.

DATASET

The dataset is looking at the top 25 news headlines on a certain day and is also giving data on how the stock market changed on that given day. The top 25 headlines are coming from Reddit, and the stock market data is coming from Yahoo Finance. The Label column gives the indication as to whether the stock market increased or decreased that day, where a 1 shows an increase and a 0 shows a decrease. A word tokenizer will be used to tokenize all of the words and make it so that the computer can read each of the words in the news headlines so that the model is able to detect if the sentiment is positive or negative. Some of the potential models that we will be using are logistic regression, random forests, gboost, and xgboost. To take a look at the data, we first looked at the distribution between the word length of the single most viewed headline on a day.

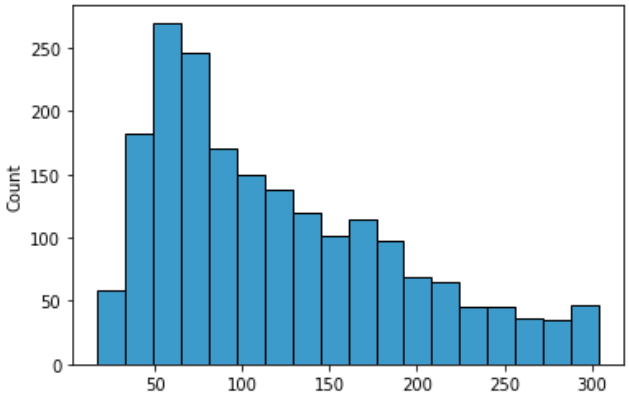


Figure 1. *News title length vs frequency*

This figure 1 shows that the number of words is typically on the lower side, but many headlines were still longer than 100 words. Here is a plot of the most common words in the dataset.

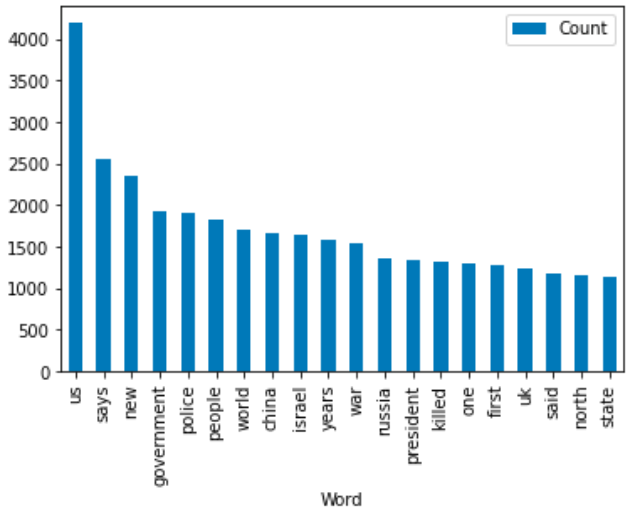
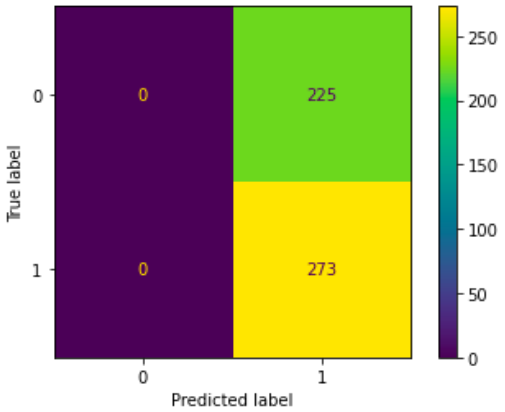


Figure 2. *Most common words from news headlines*

In figure 2, we can see that the overwhelmingly most common word was “us,” which is referring to the United States as we removed caps so that there were no repeats of words with only a difference of the capitalization. There were a little over 4000 occurrences of the word. “Says” and “new” were 2nd and 3rd respectively.

RESULTS

A numerous number of models were used. To start off, a baseline model that did not use AI was created. In this model, we calculated the mean of all of the sentiment scores, and then sorted the days with a sentiment more positive than the mean to cause an increase in the stock market, and the days with a sentiment more negative than the mean to cause a decrease in the stock market. Overall, this model was not very effective, as it had an accuracy of 50.3%. Now, to improve off of this, we decided to use a logistic regression model. Overall, this model had an accuracy of 54.8%, but to analyze results further, we must analyze the confusion matrix of the prediction.



Figer 3. *Base model confusion matrix*

From this confusion matrix, we can see that this model predicted every single day to be positive, meaning that the model said that each day, the stock market would increase in value. This is clearly inaccurate, as if no days are being predicted as negative, this program is effectively useless.

Random Forests

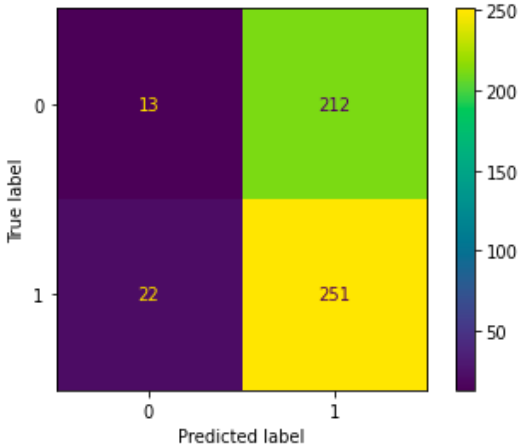


Figure 4. *Random forests confusion matrix*

To move on, the next model used was random forests. This model came out with an accuracy of 53.0%. While the overall trend is still that most days were predicted to be positive, there were a few amounts of true negative predictions in this case, which is better than the logistic regression model.

Gboost

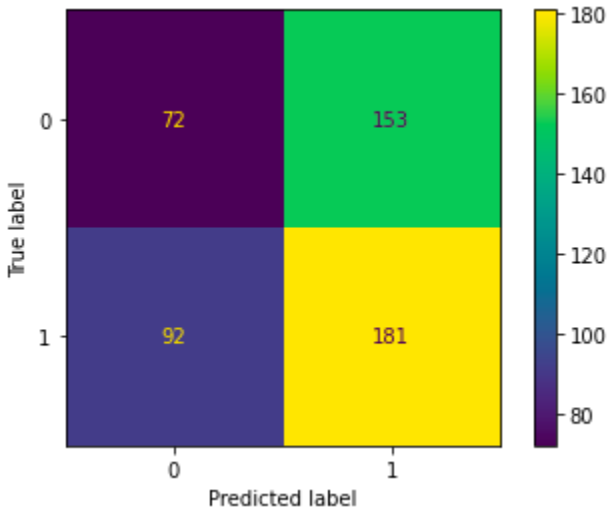


Figure 5. *GBoost confusion matrix*

GBoost was then tried, which ended up giving an accuracy score of 50.8%. This confusion matrix has a lot more false negatives and false positives, but this shows that at good amount of the days are being predicted to be negative, which is better than the logistic regression model as it is not simply predicting that every day, the market will go up.

XGBoost

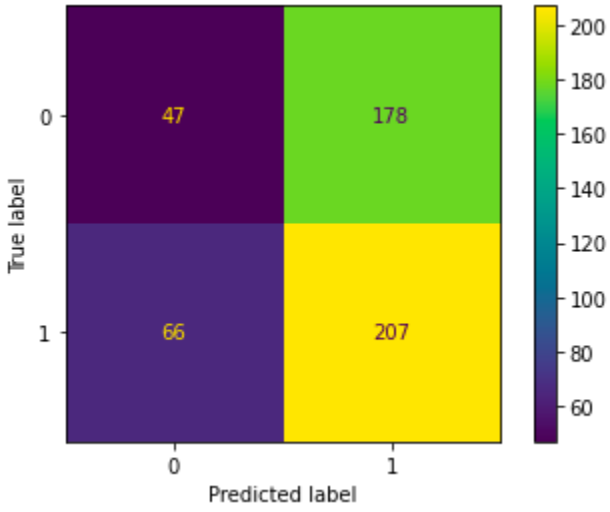


Figure 6. *XGBoost confusion matrix*

To move on from GBoost, the next model used was XGBoost. The XGBoost came out with an accuracy score of 51.0%. While this was an improvement from GBoost, it was not as major of an improvement. There were less true negatives,

which is what we are trying ot make more of to make the model to be as accurate as possible since the majority of the predictions happen to be true positives at the moment.

KNN

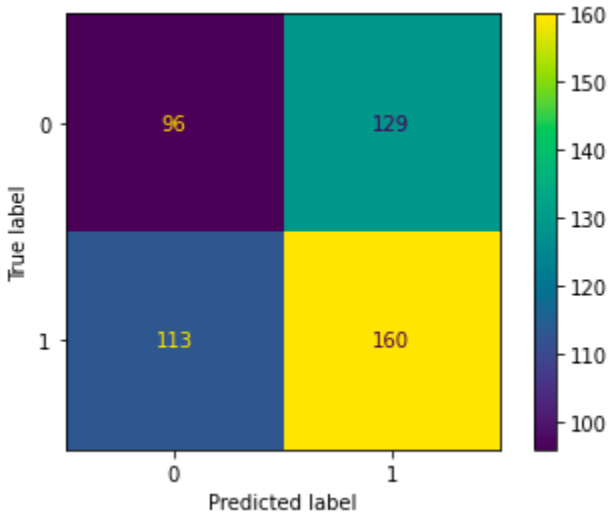


Figure 7. *KNN confusion matrix*

Next, we tried a KNN model, which gave us an accuracy score of 51.4%. There were many more true negatives than other models, but also a lot less true positives, which is not ideal, as that shows that the model was not able to successfully predict whether the market would go up nor down.

SVC

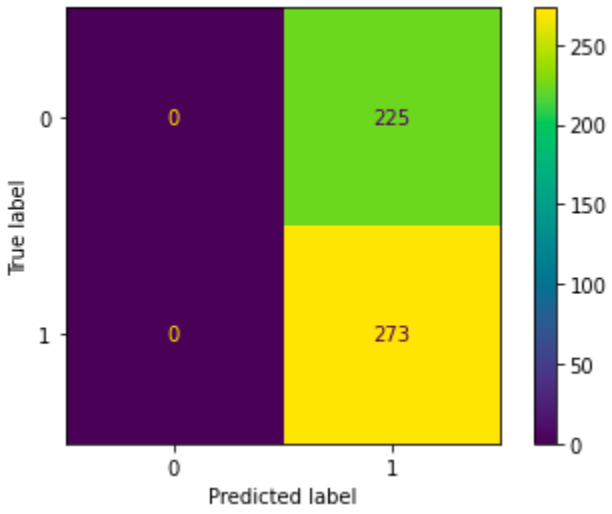


Figure 8. *SVC confusion matrix*

Next with SVC, this model was a bit more accurate, with an accuracy score of 54.8%. As we can see from this confusion matrix, this SVC model is similar to the logistic regression model in that it does not predict any day as negative. The model

predicts every result to be positive, which is not useful to us as the market is bound to have some variability and go up only about half of the time.

Gaussian Naïve Bayes

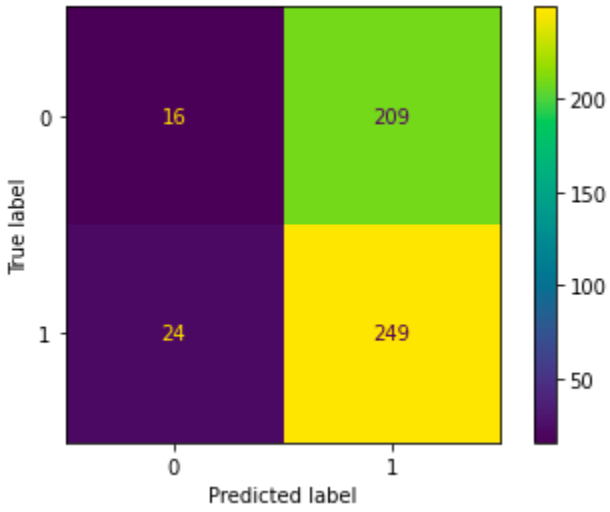


Figure 9. *Gaussian Naïve Bayes confusion matrix*

Our final model that we attempted to use was Gaussian Naïve Bayes. This model resulted with our highest accuracy score with a 55.0%. As similar to many of the models discussed, the prediction was for the most part always positive, showing that this model was not very effective, despite having the highest accuracy score of them all.

CONCLUSION

From the research conducted, it can be seen that we have not been able to find a strong relation between news headlines and whether the stock market goes up or down. No prediction model was able to give a prediction of over 60% accuracy, making guessing pretty much the same odds as if an AI model were to run, as guessing gives a 50% chance to guess the right prediction. In the many models tested, most of the models resulted in giving the result of any news article to just be positive. Unfortunately, this did not solve our problem, as if everything is predicted to cause the market to go up, there is no use in the model as they will end up being right 50% of the time and wrong 50% of the time. Though this is the case, there were many limitations with this study. One of these was the fact that we only analyzed overall market changes rather than specific changes such as a certain specified stock. Another limitation is that since we only used data from reddit, which could potentially be less credible than other sources.

REFERENCES

[1] Lee, Kari, and Ryan Timmons. Predicting the Stock Market with News Articles. nlp.stanford.edu/courses/cs224n/2007/fp/timmonsr-kylee84.pdf.