Project report

Stock market prediction (early signal model) for most trending S&P 500 stocks on Reddit

Bernd Prostmaier

August 15, 2021

Abstract

The popularity of microblogging platforms increased significantly in recent years. Based on the recent development of GameStop, it could be seen that social media sentiment can have substantial economic value. This research creates an early signal model for the three most discussed S&P 500 stocks on Reddit. Therefore, more than 35,000 comments from 4,000 submissions were evaluated. After identifying the ten most discussed stocks, further analysis provide information about their sentiment. Every comment was divided into one of three categories: positive, negative and neutral. A random forest algorithm predicts the stock price development within the next five days based on its technical indicators. The results show that in most cases the random forest prediction is consistent with the general sentiment indication.

Nevertheless, it should be stated that if no stock stands out due to its number of social media appearance, it is probably not possible to draw any conclusions about the share price performance purely on the basis of the number of mentions or the sentiment. However, in order to show truly valid evidence for statistical significance of sentiment data, sentiment scores would have to be collected on a daily basis over a period of several months.

Contents

1 Introduction

In the past the popularity of microblogging platforms increased significantly. It can be seen that financial investors are more and more interested in social media publications. One of the most popular examples in recent history is the stock price development of GameStop.

Gamestonk!! (Elon Musk, 2021)

When Elon Musk tweeted 'Gamestonk!!' with a link to the wallstreetbets Reddit thread, the stock price accelerated to more than $10bn in after-hours trading. One word or eleven characters were enough to cause the share price of GameStop to skyrocket within one day. The 'Gamestonk' post - a combination of 'GameStop' and 'stonks', which is a slang term for stocks - was liked by over 240k people and retweeted by thousands, resulting in an enormous hype about GME. This example illustrates that social media sentiment can have substantial economic value. In some cases, social media publications seem to be an indicator of whether the share price of a particular stock is more likely to increase or decrease in the near future.

2 Research objective

To get an idea of which stock can be 'the next GameStop', this research tries to identify the most mentioned stocks on social media (i.e. Reddit) and thereupon to predict the stock price development in the near future based on the stocks technical indicators. An early signal model showing buy and sell signals for these most mentioned stocks should be created. Furthermore, the theoretical foundations of the following hypotheses assume that investors do not act in a fully rational manner.

Hypothesis: The more a stock is discussed in Reddits hottest submissions, the greater its social media sentiment serves as an indicator of near-term performance.

3 Data and methodology

My project work is divided into two separate classes, each working on different tasks and can be used for separate analysis if needed. On the one hand, this research deals with data collection from the well-known Reddit forum, on the other hand, the data is imported into a machine learning algorithm predicting the near-term performance of the most discussed stocks. Both approaches are explained in more detail below.

3.1 Social media publications

3.1.1 Reddit

Founded in 2005, Reddit is one of the fastest growing social media platforms of different people sharing the things they care about most. As of January 2020, Reddit had already 52m daily active users as well as 50bn monthly views. Therefore, Reddit is one of the five most visited websites in the U.S., making wallstreetbets one of the most popular subreddits regarding online discussions about investments. In total, there are more than 100k subreddits.

Data generation took place via the python library 'praw', which allows extensive data queries via the Reddit API. This research concentrates on four subreddits: wallstreetbets,

StockMarket, Stocks & investing. To ensure that this project identifies the latest trends, data from Reddits 'hottest' submissions (i.e. posts) were analyzed. 1,000 posts from each subreddit were scraped. In total, every comment out of those 4,000 submissions was assessed individually. In particular, it was checked whether one of the S&P 500 ticker was mentioned within the comment and if so, the whole comment was saved for later use.

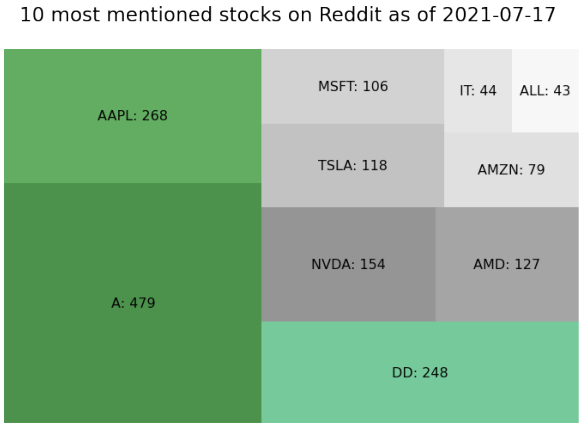

Figure 1: Ticker counting

Figure 1 shows the result of scraping those subreddits and counting the respective S&P 500 ticker. It can be seen that Agilent Technologies, Apple and DuPont were the most mentioned stocks as of 17 July 2021. This serves as a starting point for all further analyses of my project work. Based on these ten most discussed stocks, a sentiment analysis was performed for each saved comment in the next step. At the end, a machine learning algorithm will predict the future performance of the three most discussed stocks based on training data from the remaining seven stocks (see figure 1).

3.1.2 Sentiment analysis

To get an idea of whether there is good or bad sentiment, a sentiment analysis was performed for each stock in figure 1. Therefore, the so called 'vader' lexicon was used. It is a rule-based sentiment analysis tool specifically designed for social media sentiments. The vader lexicon already understands conventional use of punctuation to signal sentiment intensity (e.g. 'Good!!!') and contains a number of utf-8 encoded emoticons and slang words such as 'sux', 'kinda', etc. It is also sensitive to express both the polarity and the intensity of sentiments in social media contexts. There are already over 9,000 token features, rated on a scale from -4 (extremely negative) to 4 (extremely positive) with allowance for 0 (neutral or neither, N/A).

In order to get the best possible results from the sentiment analysis, the lexicon has been extended with a number of other 'Reddit specific' terms. Therefore, new words were assigned to the lexicon and a corresponding sentiment score was assigned. See the following examples:

- **Positive sentiment**: moon (4.0), undervalued (3.0), stonk (2.5), rocket (2.5), highs (2.0), bullish (3.0)
- **Negative sentiment**: gtfo (-4.0), hindenburg (-4.0), bear (-3.0), drop (-3.0), short (-2.0), bagholder (-1.5)

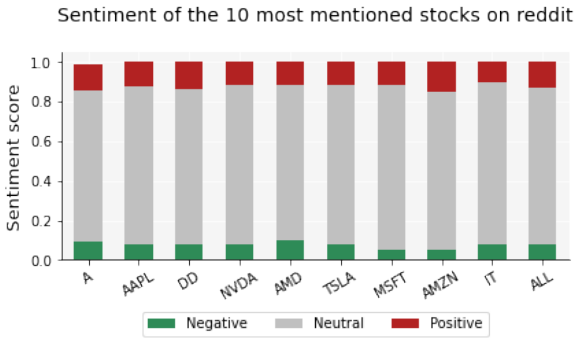

Figure 2: Sentiment scores

After extending the lexicon, the comments to be scored were cleaned up, more precisely, the text was changed to lowercase and line break characters (e.g. \n) were removed. Further

adjustments were not necessary, since no html tags are transferred from the Reddit API and emoticons are already included in the lexicon. Figure 2 shows the sentiment analysis results using vader lexicon.

However, vader not only provides positive, negative and neutral sentiment measures, but also provides a so-called 'compound' score, which is computed by summing the valence scores of each word in the lexicon and normalizes to be between -1 (most extreme negative) and +1 (most extreme positive). It could be understood as a 'normalized, weighted composite score'. According to the documentation of the vader lexicon, this is the most useful metric if one wants a single unidimensional measure of sentiment and is commonly used for sentiment analysis by most researchers. This research used standardized thresholds for classifying sentences as either positive, neutral or negative. Used threshold values are:

- **Positive sentiment**: Compound score > 0.5
- **Neutral sentiment**: $-0.5 \leq$ compound score ≤ 0.5
- **Negative sentiment**: Compound score < -0.5

Sentiment	Ticker	Comment
Positive sentiments		
positive	A	i'm very bullish on a and have been since it was oac. companies like these are the future of healthcare and telemedicine.
positive	AAPL	aapl underpriced compared to peers and ready for upside move.
positive	DD	knock the millionaire next door. should have added more dd to the moon.
Negative sentiments		
negative	A	watch a crashing. hindenburg alert
negative	AAPL	aapl gets slapped!!!!
negative	DD	dd really? ohhhh boy please nooo
Neutral sentiments		
neutral	A	early a was the best a
neutral	AAPL	anyone here thinking opex might be priced in for aapl and this thing could fly?
neutral	DD	one dd and suddenly ketchup hype?

Figure 3: Sentiment classification

Figure 3 shows some examples for classification based on the previously mentioned compound score. The compound score will serve as a basis for the later reconciliation of the results from the machine learning algorithm. Recall, the hypothesis says that the most discussed stocks will tend to rise/fall in the next few days according their positive/negative sentiment. Therefore, an overall positive sentiment would tend to indicate rising stock prices and an overall negative sentiment would tend to indicate falling stock prices.

3.2 Machine learning

The second part of this project deals with the prediction of the stocks closing price in the next days. To be more specific, it was examined whether the share price of a particular stock will be higher or lower in five days than it is today.

3.2.1 Random forest

As this will be a classification task, either -1 (stock price will decrease in five days) or +1 (stock price will increase in five days), random forest classification was used as a machine learning method to predict the stock price development. Random forests provide an improvement over simple classification trees and bagged trees by way of a small tweak that

decorrelates the trees. A number of decision trees B is built on bootstrapped training samples, but at each time a split in a tree is considered, a random sample of m features (out of all p features) is chosen as split candidates. On average $(p-m)/p$ of the splits don't consider one strong feature and so other features will have more of a chance. This results in decorrelating the trees, thereby making the average of the resulting trees less variable and hence more reliable. In other words random forests calculate $\hat{f}^1(x), \hat{f}^2(x), ..., \hat{f}^B(x)$ trees using B separate training sets and average them in order to obtain one single low-variance random forest model given by:

$$\hat{f}_{avg}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^b(x)$$

Consequently, all of the bootstrapped trees B don't look as similar to each other as they probably would if we chose bagging. Averaging those low correlated quantities will lead to a reduction in variance.

3.2.2 Model preparation

Since the sentiment data is only available for a specific day, technical indicators of the respective stocks were used as features on a daily basis for the last 18 months. Of course sentiment data can be used as a feature as well, but this would require running the sentiment analysis for several months first while storing the sentiment results for each day. Stock market data since 1 January 2020 were fetched via the 'yfinance' library. The 'TA-Lib' library was used to perform technical analysis in order to compute technical indicators such as the relative strength index, rate of change, on-balance volume, etc. on a daily basis. The derived technical indicators, in combination with the stock market data from yfinance, formed the feature set for the random forest algorithm.

Since this research is interested in the price development of the three most mentioned stocks (see figure 1), the remaining seven stocks serve as training-/test data. Therefore, the model can be fitted with more than 2,400 samples. In order to optimize the hyperparameters for the random forest model, 10-fold cross validation was used to fit 1,000 different models. The main parameters, varied during the cross validation, were:

- **Number of estimators**: Number of trees in the forest (range: $10 - 1,000$)
- **Min samples leaf**: Minimum number of samples required to be at a leaf node (range: $1 - 200$)
- **Max depth**: Maximum depth of the tree (range: $50 - \infty$)
- **Min samples split**: Minimum number of samples to split an internal node (range: $2 - 200$)
- **Max features**: Number of features to consider when looking for the best split (auto, $\sqrt{m}$, $log2(m)$, None)

Since our label is a simple binary value, the 10-fold cross validation error is given by the average number of miss-classified samples:

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

The result of validating over 1,000 models using 10-fold cross validation stated that the random forest model using the following set of hyperparameter results in the lowest classification error: Number of estimators = 750, Max depth = 1,500, Min samples leaf = 1, Min samples split = 2, max features = 'log2'.

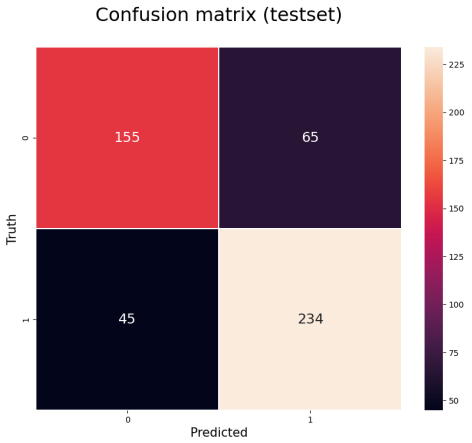

Figure 4: Testset confusion matrix

Cross validation stated that the random forest model should consider $m = log2(p)$ features at each split in the tree in order to obtain the best possible fit, which corresponds to the described advantage of random forests (decorrelation) described in chapter 3.2.1. Using this set of hyperparameter determining the predictive model, the random forest predicts the stock price development in the testset with an accuracy of 78.0%! Figure 4 shows the training data confusion matrix.

3.2.3 Prediction & early signal model

In the last step the predictive model was used to predict the stock price development of the three most discussed stocks from the Reddit analysis in section 3.1 (unseen data).

It can be seen that the random forest model predicts the stock price development within the next five days according to the overall sentiment score (compound) from section 3.2. However, to be able to make statistically significant statements, this analysis would have to be carried out over a period of months. This would allow to see whether there is truly some form of dependence.

Ticker	Sentiment	Random forest prediction*
Top three discussed stocks from Reddit		
A	positive	1
AAPL	positive	1
DD	neutral	-1

** 1 = Increasing stock price; -1 = decreasing stock price*

Figure 5: Reconciliation: Sentiment vs. machine learning results

In order to give the project a long-term character, an early signal model was implemented based on the machine learning results. The specially developed algorithm checks every five days whether a reversal of the price trend is taking place. If this is the case, the algorithm sets buy or sell signals. Figure 6 shows the stock performance of Agilent Technologies including the determined buy and sell signals over the last 18 months.

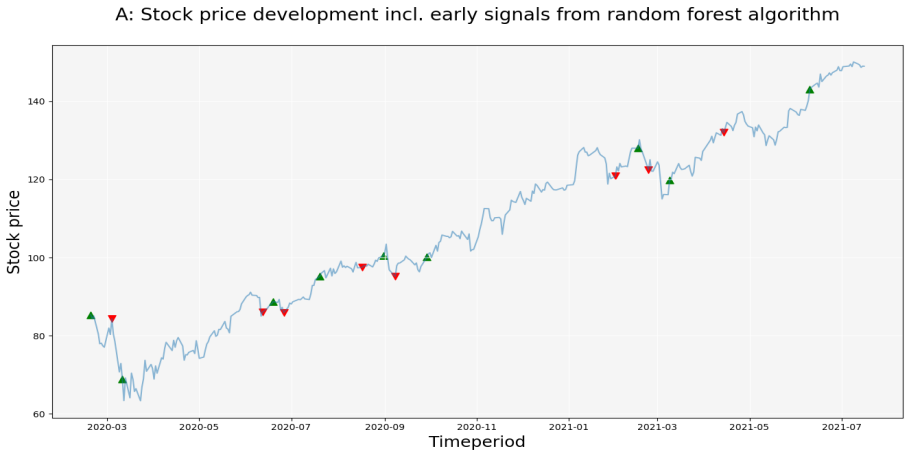

Figure 6: Early signal model for Agilent Technologies

4 Results and conclusion

This project shows that social media sentiment is consistent with predicted price trends based on the stocks technical indicators. During the research, the correspondence of the sentiment scores with the predicted price movements were monitored on 10 days. In most cases, at least two of the three stock predictions matched with their sentiment indication. However, to be able to make a valid statement, this research would have to be carried out over a longer period of time. In order to show truly valid evidence for the statistical significance of sentiment data, the sentiment scores would have to be collected on a daily basis over a period of several months. This would also allow the sentiment score to be included as a feature variable. However, for the time frame of this project, that would be impossible.

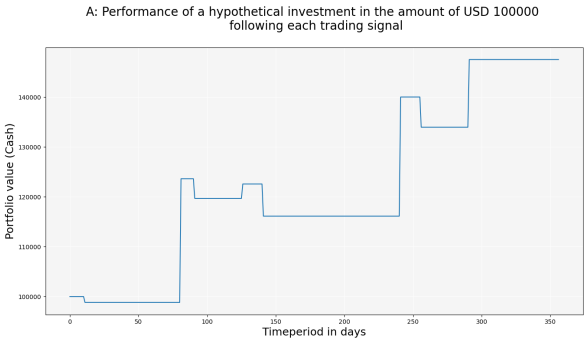

Figure 7: Hypothetical investment

Nevertheless, the GameStop share price development mentioned at the beginning of this paper is likely to have been a special case. The analysis of sentiment data during the period of this research did not show any particular anomaly. No stock stood out in particular, e.g. no stock was mentioned 10x or 100x more often than its following most mentioned stock on any given day. All in all, it can be said that the number of discussions about a certain stock or its sentiment on social media can be a good first indication for stock picking. However, it does not replace a careful analysis of the respective stock using more conventional methods. The two coded python classes of this project provide valuable information about the most mentioned stocks on Reddit as well as an early signal model with corresponding plausible prediction results for those stocks. Figure 7 shows the development of a hypothetical investment in the amount of USD100.0k following each trading signal in figure 6. Overall, the model provides satisfactory results, meeting the research objectives stated in chapter 2 of this report.

References

[1] Shead, S. *Elon Musk's tweets are moving markets — and some investors are worried, 2021, https://www.cnbc.com/2021/01/29/elon-musks-tweets-are-moving-markets.html*

[2] Reddit. *Dive into anything, 2021, https://www.redditinc.com/*

[3] cjhutto. *vaderSentiment, 2021, https://github.com/cjhutto/vaderSentiment*

[4] Hastie, T., James, G., Tibshirani, R., Witten, D. *An Introduction to Statistical Learning, 2013*