UNIVERSITY OF MALAYA

Recommending potential stocks for investment

WQD7005 Data Mining

Lecturer: Dr. Teh Ying Wah

Name	Matric Number
Noraisha Yusuf	WQD 180008
Group Members	
Marina Shah Muhammad Zabri Tan	WQD 180011
Amira An-Nur binti Rusli	WQD 180016
Aminah Sofia bt Mahayudin	WQD 180032
Norbaizura Mohamad	WQD 180043

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

UNIVERSITY OF MALAYA

2019

Contents

1 Objective

Stock investment is one of the highly researched area and hot topic for analysis. There have been various theories and techniques used to study and analyze stocks, influential factors and effects. Further, there are divergent views on suitable investment strategies. The purpose of this project is to identify and integrate multi-view dataset, through which machine learning techniques can be applied to recommend potential stocks for investment.

2 Dataset

As part of the group milestones, we crawled 3 months' worth of stock data from Market Watch portal on www.thestar.com.my. We crawled opening, closing, high and low prices as well as trade volume for all stocks (recorded on the portal) from 1st January 2019 to 1st April 2019. From this web crawl exercise, we accumulated information for 1805 stocks, with a total of 71,222 data points. Additionally, we crawled the latest quarter financial information from the i3investor portal. News title and introduction paragraph is crawled from The Star Online. Tweets were crawled using 'Twitterscraper' package. Finally, data is downloaded from Google trends which enumerates and scaled the number of searches of the stock companies. Table 3.1 shows the summary of 5 types of dataset used in this project.

Table 3.1: Summary of dataset collection

No.	Dataset	Description
1	Stocks	Crawled opening, closing, high and low prices as well as trade volume *Source: The Star online, Market Watch*
2	Financial	Crawled PBT, EPS, Revenue, dividend and many more *Source: https://klse.i3investor.com/financial/quarter/latest.jsp*
3	News	Crawled News title and first paragraph of news article *Source: The Star online*
4	Tweets	Crawled Twitter messages, date and author *Source: Twitter*
5	Google Trends	Search queries relating to company on Google. This is the only data not crawled, and downloaded directly from source in CSV format *Source: Google Trends*

3 Schema

Based on the collection of datasets, a data warehouse schema is developed as shown in Figure 4.1.

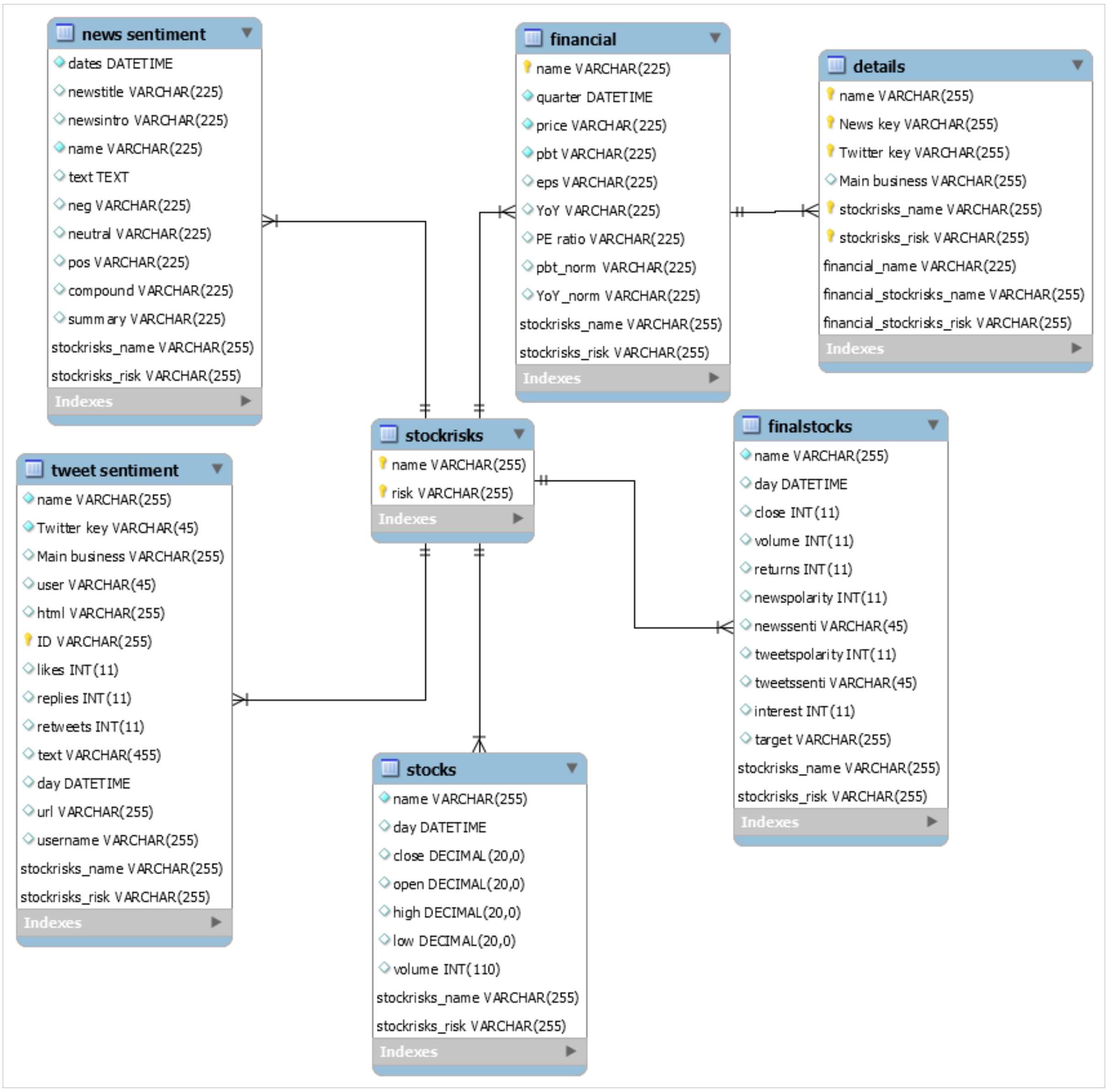


Figure 5.1: Snowflake schema

4 Methodology

4.1 Diversification: risk vs. return

The rule of thumb in investment is diversification of portfolio, to include multiple investments in one portfolio. If one investment loses money, the other investments may compensate or yield higher return. There are divergent views on this theory. According to the reputable American investor Warren Buffett, he has famously stated that "diversification is protection against ignorance. It makes little sense if you know what you are doing.". Basically, he is against the theory of diversification and that one will benefit more by running a concentrated portfolio. On the other hand, Dr. Michael Burry[1] recommended investors to develop their own investment style, and that diversification is not necessarily a strategy of the ignorance and lazy as claimed by Warren Buffet. Traditional portfolio theory states that maximizing the number of investments to the maximum limit could minimize the risk of the overall portfolio. Nonetheless, according to the empirical research by Wang (2010), he recommended that between 20 to 30 stocks to achieve optimal portfolio size, in which the diversification risks is minimized whilst not offsetting the benefits of diversification i.e. returns of the portfolio is not extensively sacrificed. The fundamental idea behind portfolio diversification is to be able to strategize the choices of stocks, by achieving a trade-off between risk and reward (i.e. returns) of the stocks. To construct a diversified portfolio, a common method is to identify common movement of stocks based on correlation matrix of stocks (Eom & Park, 2017). In view of this, this project will explore clustering method to identify clusters of correlated stocks and subsequently map the clusters based on stock returns vs. volatility plot to determine the high, medium and low risk clusters of stocks. Figure 4.1 shows the process flow to identify three clusters of stocks that have varying risk-return outcome. Stocks can then be chosen from each of the clusters, thus diversifying our portfolio.

[1] An American physician, investor, and hedge fund manager who are among the few ones making significant returns by betting against the subprime mortgages, during the subprime crisis in 2009.

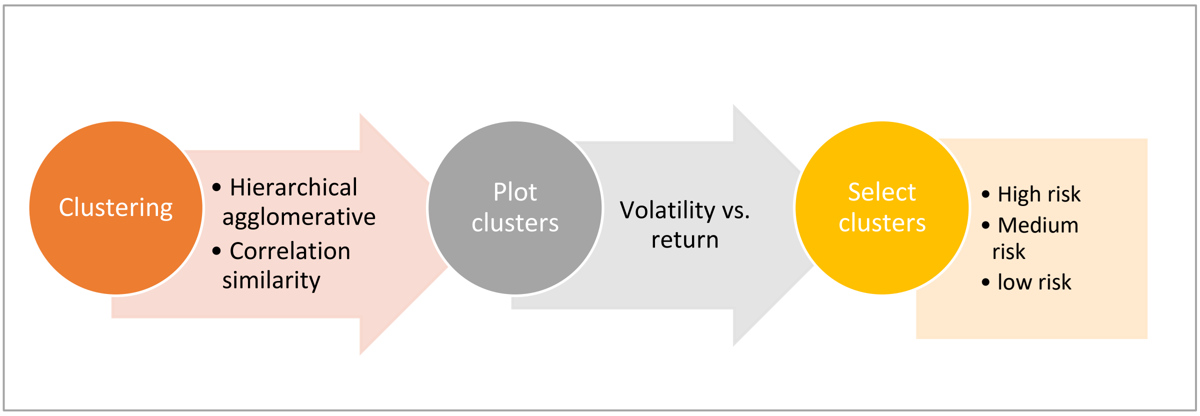


Figure 4.1: Process flow to determine stock risk

Quarterly average of volatility and return of stock is computed.

- Stock returns is computed based on the change in daily stock prices is computed. Specifically, logarithmic return formula is used. The advantage of using log differences is that this difference can be interpreted as the percentage change in a stock but does not depend on the denominator of a fraction. Averaged value is then computed and multiply with number of days stock market is in operations in a quarter i.e. 63 days

$$Averagestockreturn = \frac{\sum log(closedprice_i) - log(closedprice_{i-1})}{no.ofdays} \times 63days$$

- Volatility is a statistical measure of the dispersion of returns for a given security or market index. In most cases, the higher the volatility, the riskier the stock. Standard deviation of the daily closing prices is computed and multiplied with number of days stock market is in operations in a quarter i.e. 63 days.

$$Volatilitystockreturn = \sqrt{\frac{\sum(dailyreturn - mean(return))^2}{no.ofdays - 1}} \times \sqrt{63days}$$

4.2 P/E ratio: overvalued stocks

Having identified the different risk categories of stocks, the next step is to figure out which of the stocks are worth investing in. Price to Earnings (P/E) ratio is the most widely used as a factor in investing. It reflects the amount that investors are willing to pay for the stock in return of earnings. The higher the P/E ratio, the higher investor's expectations of the earnings.

However, high P/E ratio may also indicate overvalued stock (Andy, 2018), whereby the stock is priced significantly higher relative to its earnings e.g. Stock price is $100 and earning per share is $2 which gives a P/E ratio of 50. In the article by Andy (2018), he recommended that investors make comparison between the P/E ratio and the company's growth. Based on this, the stocks will be mapped out and P/E ratio of stocks will be compared against the company's profit before tax and year-on-year growth. This is to select potential valuable stocks.

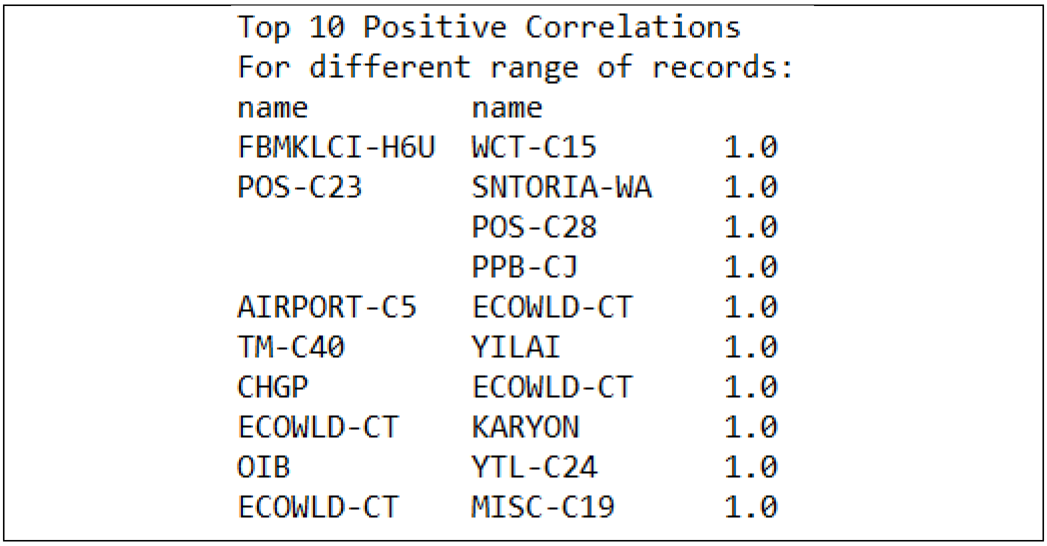
4.3 Sentiment and interests in company

Various studies have suggested that market sentiment can influence stock prices. Stocks with higher returns are more prone to be affected by market sentiments. In particular, negative market sentiment has a significant effect on stock prices (Allen, McAleer, & Singh, 2019). This is because investors fear of losing money, more so than having confidence of an upside gain (Dash & Maitra, 2018). Investors would act on immediately and sell their positions if downside losses is expected. To better understand sentiment influences, this project will explore news and twitter sentiment on the company and how they affect the stock prices. Additionally, the interests on the company will be measured based on number of online searches via Google trends data.

5 Results

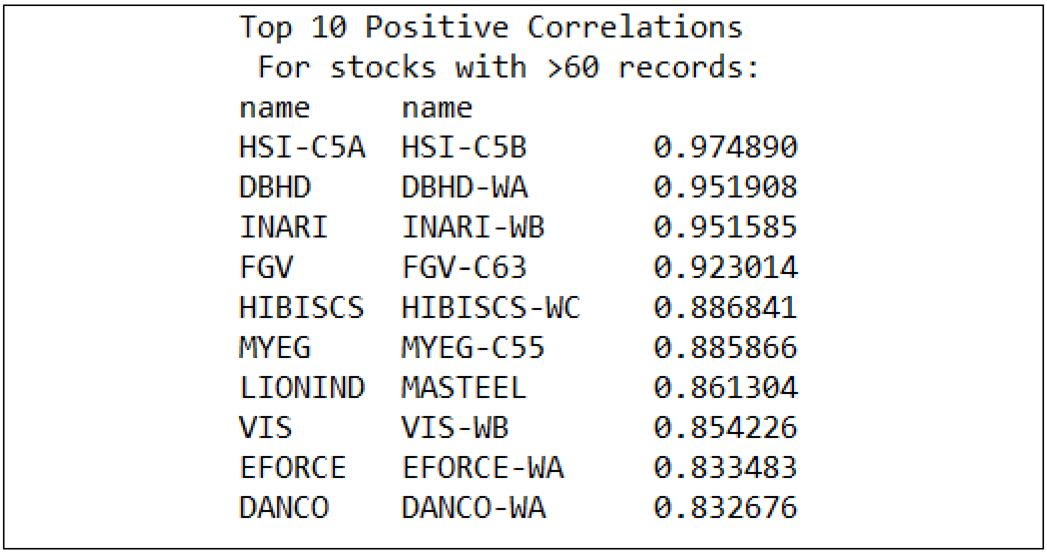
5.1 Correlation of stocks

Initially, information for 1805 stocks were crawled for a period of 3 months. Upon inspection, some of the stocks were only in the market for several days only while others are trading fully throughout the three months. The mismatched of data points can lead to bias correlation values because correlation is supposed to be computed for stocks with equal trading period. Figure 5.1 demonstrates this bias, whereby the top positive correlation values achieves perfect correlation which is illogical considering how most stocks are inherently unique. Figure 5.1 shows the true correlation values, where stocks from similar trading period were used. In view of this, only stock data that contains 61 data points i.e. closing prices for 61 trading days are selected. This exercise reduces the amount of stocks from 1805 to 402 stocks.



```
Top 10 Positive Correlations
For different range of records:
name         name
FBMKLCI-H6U  WCT-C15       1.0
POS-C23      SNTORIA-WA    1.0
             POS-C28       1.0
             PPB-CJ        1.0
AIRPORT-C5   ECOWLD-CT     1.0
TM-C40       YILAI         1.0
CHGP         ECOWLD-CT     1.0
ECOWLD-CT    KARYON        1.0
OIB          YTL-C24       1.0
ECOWLD-CT    MISC-C19      1.0
```

Figure 5.1: Top 10 positive correlation values for all stock data



```
Top 10 Positive Correlations
 For stocks with >60 records:
name      name
HSI-C5A   HSI-C5B       0.974890
DBHD      DBHD-WA       0.951908
INARI     INARI-WB      0.951585
FGV       FGV-C63       0.923014
HIBISCS   HIBISCS-WC    0.886841
MYEG      MYEG-C55      0.885866
LIONIND   MASTEEL       0.861304
VIS       VIS-WB        0.854226
EFORCE    EFORCE-WA     0.833483
DANCO     DANCO-WA      0.832676
```

Figure 5.2: Top 10 positive correlation for stocks trading for at least 60 days

5.2 Clustering stock risk

Hierarchical agglomerative clustering (HAC) was performed on the filtered 402 stocks. The correlation matrix values is used as the similarity metric. 5 clusters were deemed as the best cutting point as shown in the Dendogram in Figure 5.3. The return and volatility of stock were computed based on the formula stated in section 4.1. The clusters obtain through HAC are then mapped onto the return and volatility values. Regression line is also fitted for each cluster to view the projection/trend line of respective clusters. As shown in figure 5.4, it can be observed that cluster 3 is the high-risk cluster whereby the stock returns increases sharply as volatility increases. Cluster 1 is the low-risk cluster whereby the stock returns is relatively lower as compared with other clusters. Cluster 5 is deemed as medium risk cluster whereby moderate returns are expected as volatility increases. Hence, cluster 1 (61 stocks, cluster 3 (185 stocks) and cluster 5 (78 stocks) were chosen for further analysis.

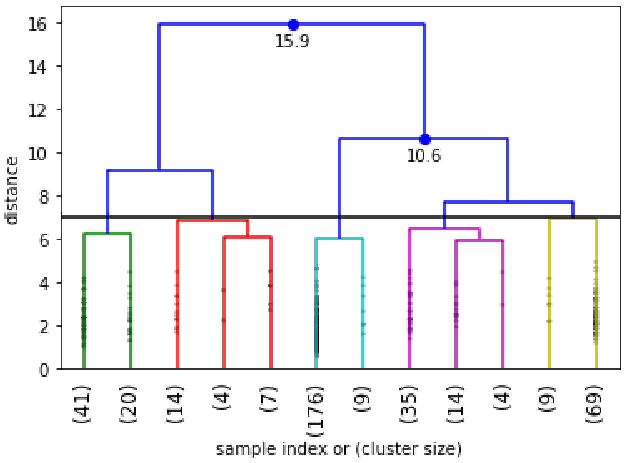


Figure 5.3: Hierarchical Clustering Dendogram

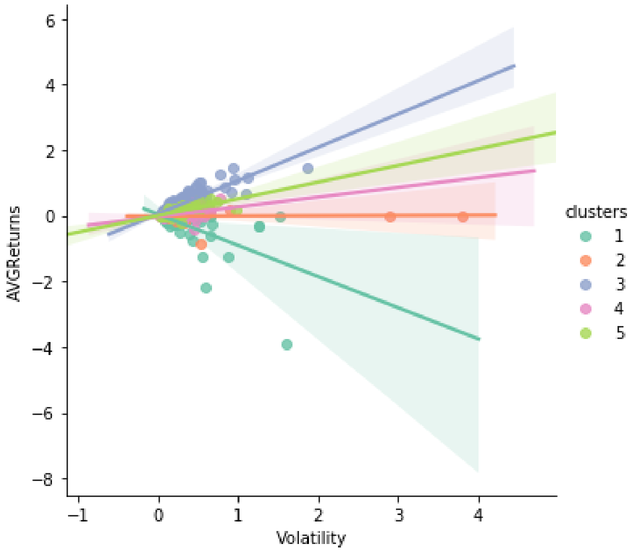


Figure 5.4: Returns vs. Volatility of stock clusters

5.3 Determine stock value

Latest quarter financial information of the companies was crawled. However, this does not contain information for all stocks. Further filtration had to be conducted. The low risk cluster reduces from 61 to 30 stocks, medium risk cluster reduces from 78 to 33 stocks and high-risk cluster reduces from 185 stocks to 80 stocks. The P/E ratio was computed for all three clusters by dividing stock prices with EPS (earning price per share). This is then compared with the year-on-year growth (YoY) and the profit before tax (pbt) of the company. Tableau tool was used to visualize this comparison (Figure 5.5). It can be observed that, most of the stocks lies between the 0 and 1.5 for PE ratio with less than 1% for YoY growth. The profit ranges with most stocks fall below RM3 Million. The key step is to filter out undervalued and overvalued stocks. In doing so, we can be ensured that we are investing in valuable or promising stocks. For this, the rules-based filtration is carried out, whereby only stocks with positive profit rate, growth rate of at least -1 % and P/E ratio of between 0 and 3 were chosen. At this stage the number of potential stocks is 46, whereby 11 are low risk stocks, 7 are medium risk stocks and 32 are high risk stocks (Figure 5.6).

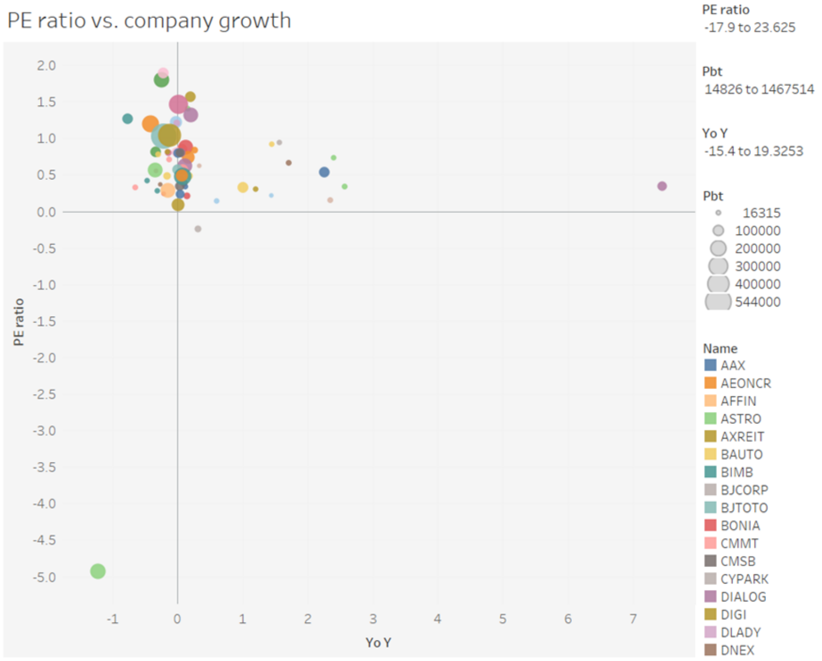
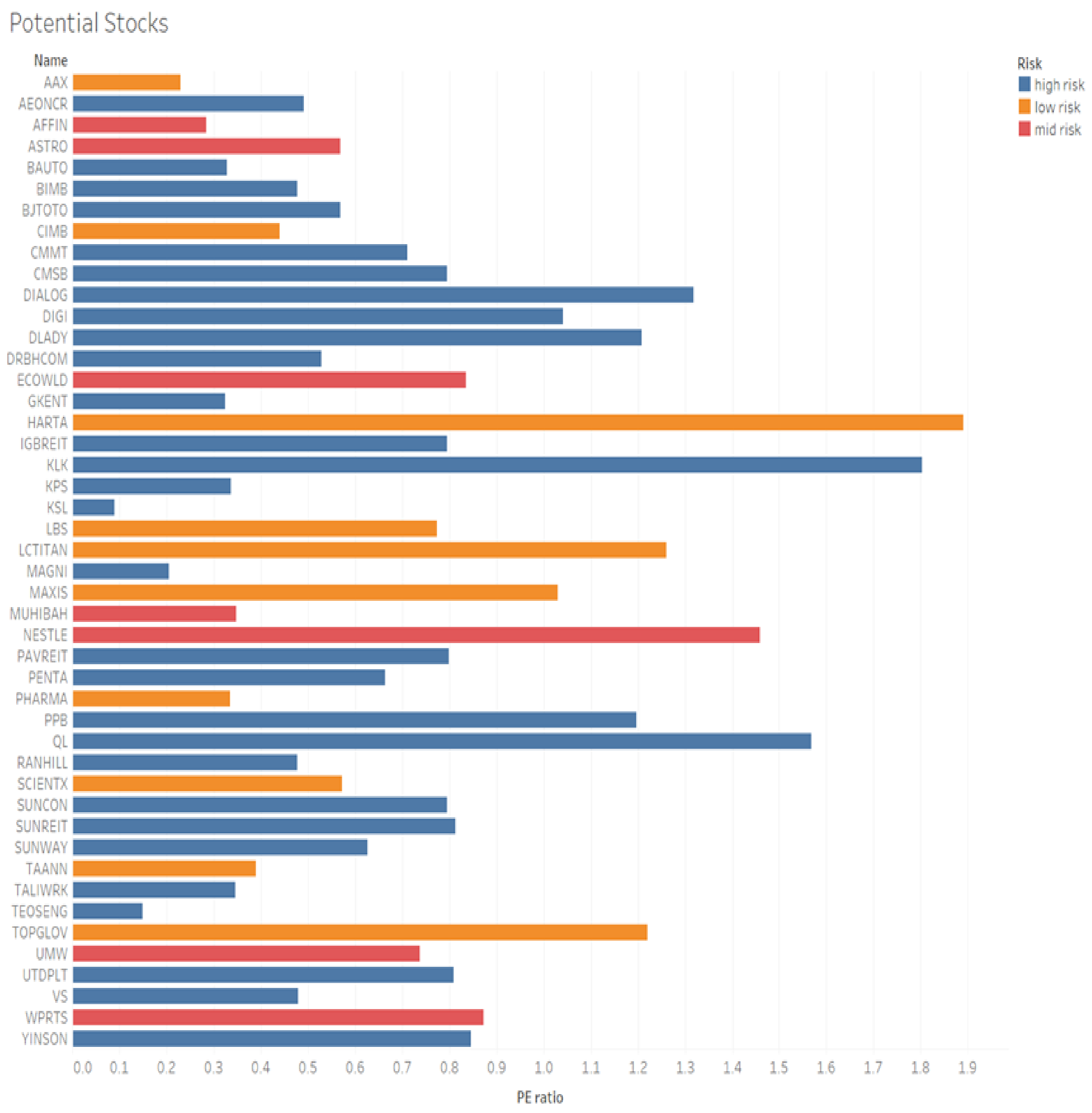


Figure 5.5: PE ratio vs. company growth and profit



PE ratio as an attribute for each Name. Color shows details about Risk. The data is filtered on PE ratio, Pbt Norm and YoY norm. The PE ratio filter ranges from 0 to 3. The Pbt Norm filter ranges from 0 to 6.006291215. The YoY norm filter ranges from -1 to 19.3253.

Figure 5.6: List of potential stocks of different risk categories

5.4 Sentiment analysis

Investors have the option of trading from this group of potential stocks and selecting from each risk category in order to diversify their investment portfolio. The next part of this project will focus on how sentiments can influence stock prices. News reported on the company on The Star and conversations about the company in Twitter were extracted. By simply looking at the word cloud (Figure 5.7 and 5.8) i.e. frequency of words of the news and tweets are not providing a lot of insights. This is due to the presence of common words. In view of this, sentiment analysis is explored.

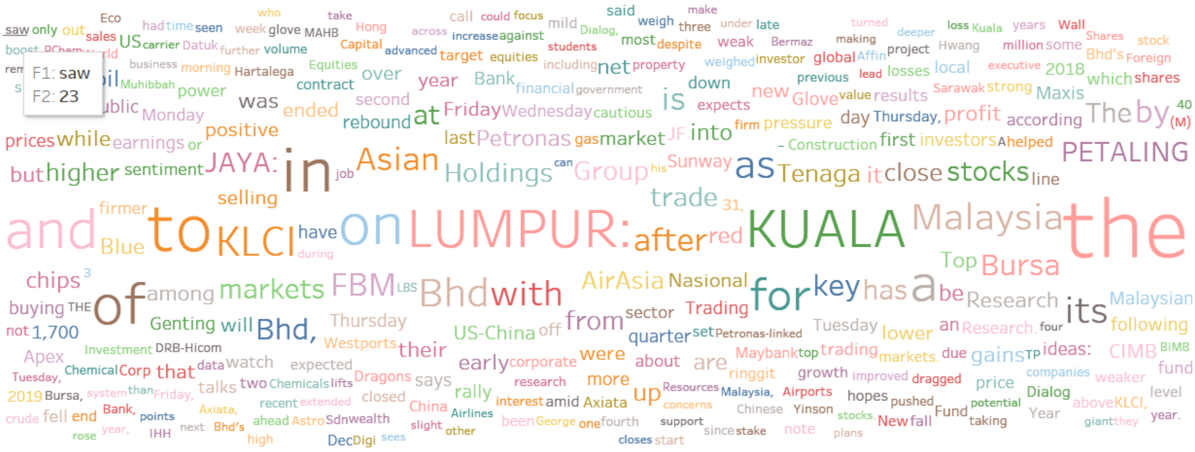


Figure 5.7: Word cloud news reports on stocks

for
that
of
from
in
a
the
to
is
with
and
you

Figure 5.8: Word cloud tweets reports on stocks

Vader, a sentiment analysis package in Python is used to perform the sentiment analysis of the news and twitter dataset. Through quantification of the textual data, sentiment analysis enables easier comparison between the news reports and tweets. When comparing the sentiment analysis results of news (Figure 5.9) and tweets (Figure 5.10), it can be observed that there are more tweet contents about the stocks as compared with news reports. Overall, there are more positive tweets than negative and neutral tweets. We need to zoom into specific stocks to better analyze the influence of sentiments on the stock prices.

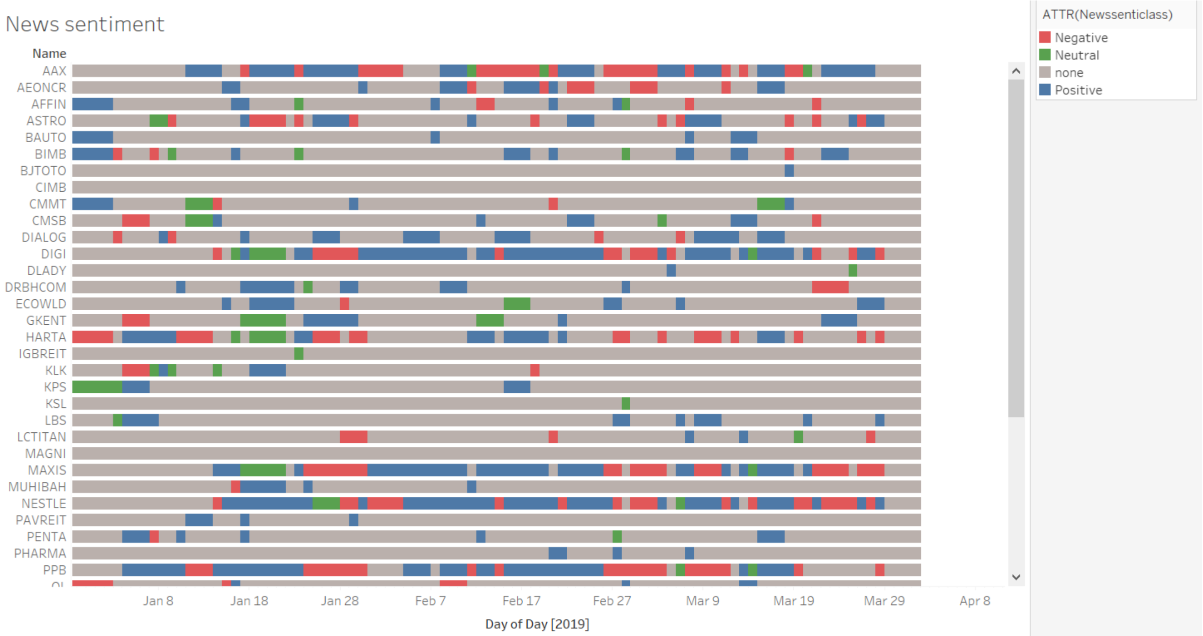


Figure 5.9: Visualizing sentiment analysis of news data

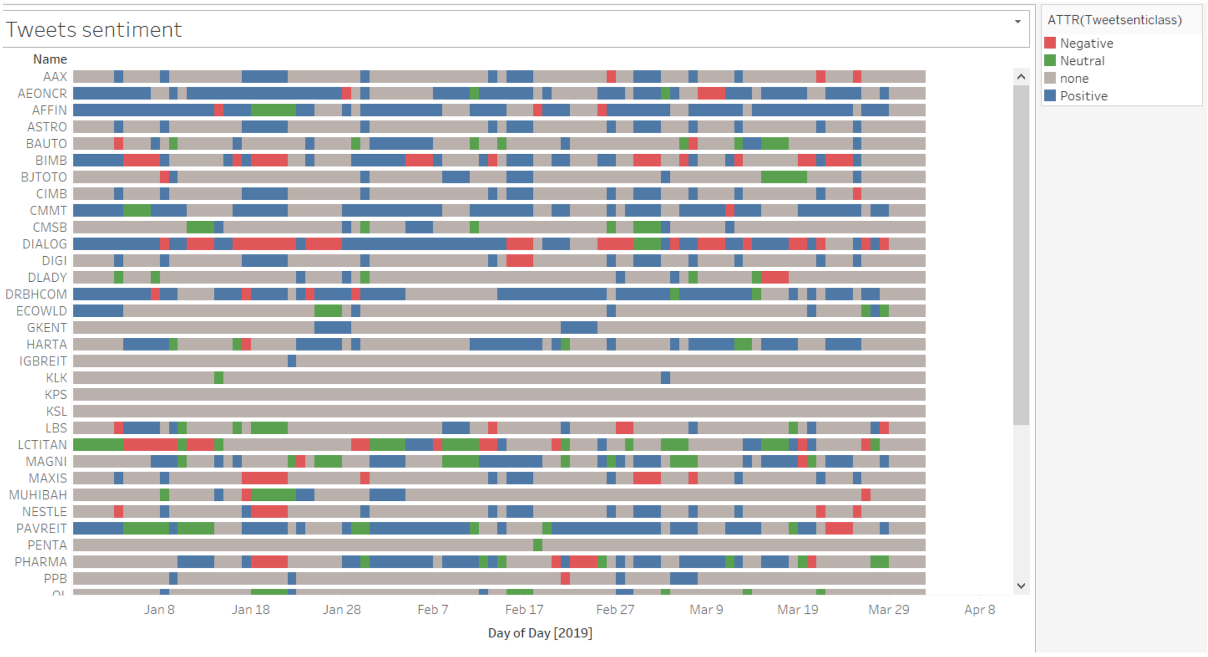


Figure 5.10: Visualizing sentiment analysis of tweets data

5.5 1-D SAX time series

Four stocks were chosen from the list of potential stocks identified in section 5.3, namely AAX, MAXIS, NESTLE and AEONCR in view that they comprise many news and tweets during the 3 months period as well as representing the different risk categories and business sectors.

Table 5.1: Stock details

Stock name	Company	Business sector	Risk category
AAX	AirAsia X Bhd	Aviation	Low risk
MAXIS	MAXIS Bhd	Telco	Low risk
NESTLE	NESTLE (Malaysia) Bhd	Consumer	Medium risk
AEONCR	Aeon Credit Service (M) Bhd	Finance	High risk

1d-SAX is a method to represent a time series of the stock returns. SAX (Symbolic Aggregate approXimation) is one of the main symbolization techniques for time series. 1d-SAX incorporates the trend i.e. slope between data points. Comparison of the 1d-SAX representation of all 4 stocks (Figure 5.11) shows that the trends and fluctuations between the stocks differs between one another. This indicates that these stocks behaved differently and possibly independent of each other. This further strengthen the diversification theory, of investing in stocks that are not co-moving, thus reducing the overall risk of the portfolio.

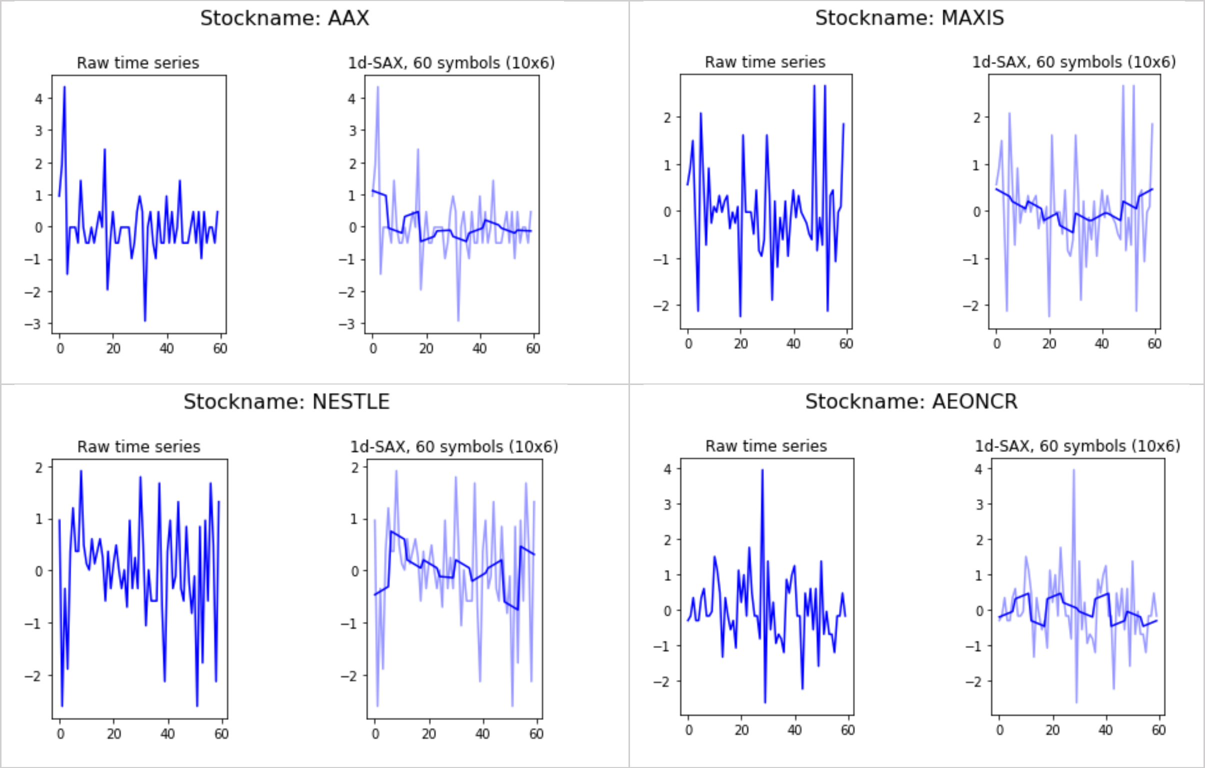
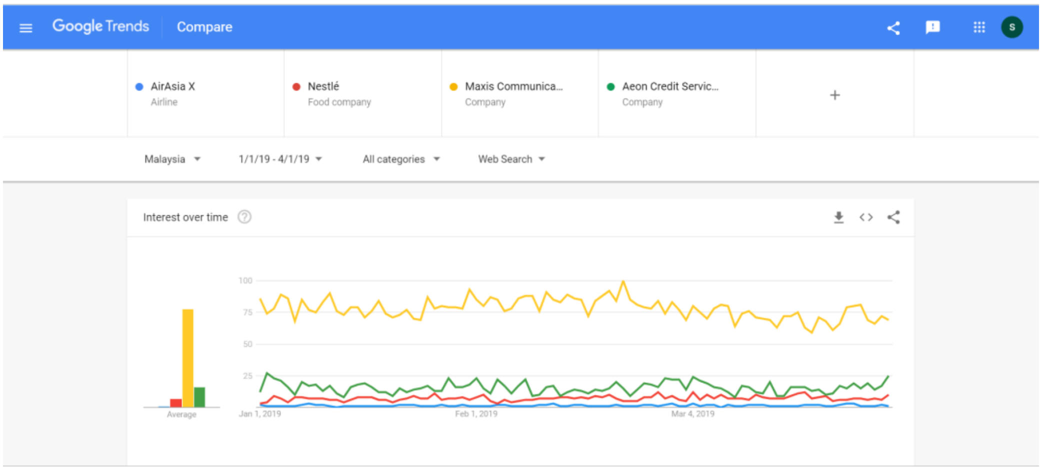


Figure 5.11: Time series and 1D-SAX dimension reduction

5.6 Google Trends data, interests in stocks

In addition to sentiment analysis, demand for the company’s services plays a role to determine whether to invest in a stock. Google Trends allows us to view the number of times the specific word is searched on Google (Figure 5.12). The results can be a proxy of the current demand for the company’s services. This is an experimental approach, to investigate whether a significant increase in the volume searches is related to an increase or decrease in the value of stocks.



AirAsia X
Airline

Nestlé
Food company

Maxis Communica...
Company

Aeon Credit Servic...
Company

Malaysia ▾ 1/1/19 - 4/1/19 ▾ All categories ▾ Web Search ▾

Figure 5.12: Google Trends search

In Figure 5.13, Nestle’s stock returns is compared against the interest value (i.e. Google Trends data). It can be observed that, the spike in the interest seems to be aligned with significant fluctuations in the stock returns. Thus, it is worth exploring the relationship between Google trends and stock.

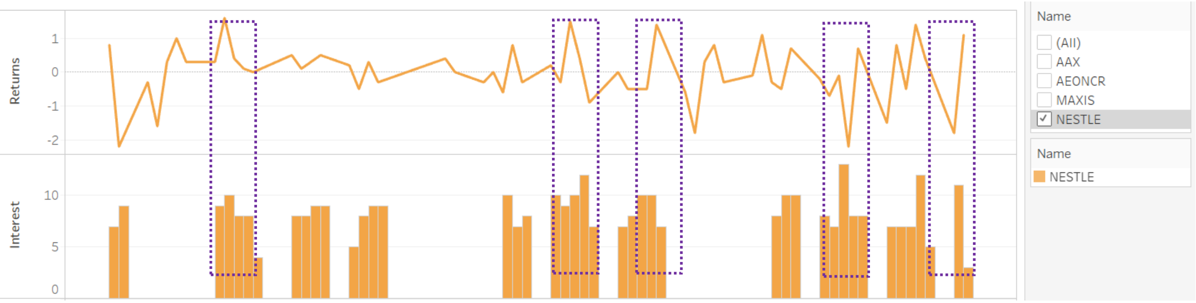


Figure 5.13: Interest, sentiments and stock returns

5.7 Machine Learning classification

For the machine learning classification model, target variable is required. Notwithstanding, the task of creating target variable for our time-series stock data is not straightforward. Initially, the idea was to create buy/don't buy classes based on the stock prices. However, this would lead to bias classification model. Further, determining the specific date to purchase is very subjective.

Thus, in ensuring a fair model, the machine learning classification model will explore the detection of changes in stock returns based on sentiments and interests. The target variable is created based on the changes in stock returns. Three classes were chosen namely down (decreasing returns), up (increasing returns) and none (no changes). The variables used for this model are sentiment scores for news, sentiment category for news, sentiment scores for tweets, sentiment category for tweets and interests (Google trends data). The stock data is excluded to reduce bias in the model. This in view that stock returns are correlated with stock data prices and trade volume.

In view that there are 3 classes, logistic regression would not be ideal. Hence, decision tree algorithm was chosen. Two decision trees were developed and compared (Figure 5.14)

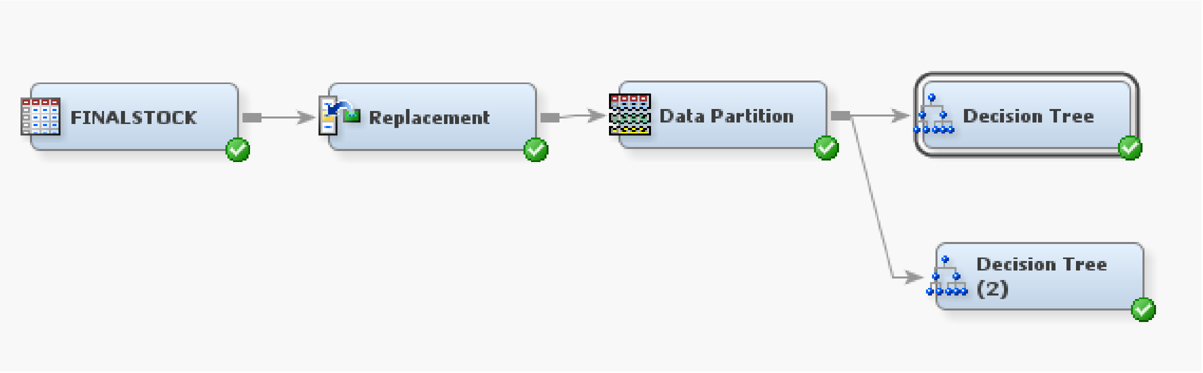
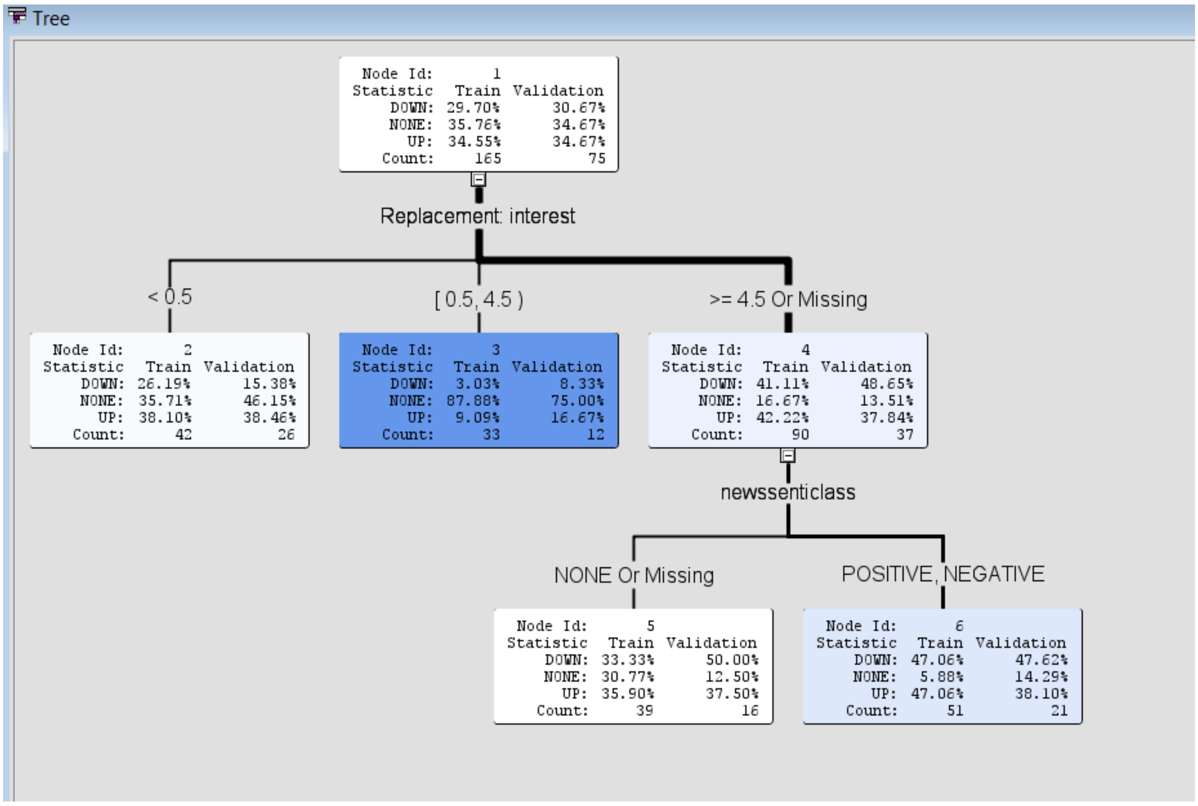


Figure 5.14: SAS Analytic workflow

The first decision tree (Figure 5.15) was generated based on probability Chi Square while the second decision tree (Figure 5.16) was generated based on Entropy with the best pruned tree selected according to average square error. Interestingly, in both trees, only news sentiment and interest variables were considered to split the tree. This could imply that twitter sentiment is not a good predictor of stock prices. This is also evident based on the assessment of variable importance (Figure 5.17). This further strengthens our findings in section 5.16, whereby interests in the form of Google search relates to stock prices and stock returns. The second decision tree outperforms the first decision tree with better accuracy of event classification rate, for both training and validation set (Table 5.2). From the same result, it is worth noting that, news sentiment and interest are both good predictors of decreasing stock returns but not for increasing stock returns e.g. the second decision tree accurately classify true negatives by more than 70% but only achieve 50% accuracy for true positives classes.



```
Node Id:        1
Statistic   Train  Validation
    DOWN:  29.70%      30.67%
    NONE:  35.76%      34.67%
      UP:  34.55%      34.67%
   Count:    165          75
```

```
Node Id:        2
Statistic   Train  Validation
    DOWN:  26.19%      15.38%
    NONE:  35.71%      46.15%
      UP:  38.10%      38.46%
   Count:     42          26
```

```
Node Id:        3
Statistic   Train  Validation
    DOWN:   3.03%       8.33%
    NONE:  87.88%      75.00%
      UP:   9.09%      16.67%
   Count:     33          12
```

```
Node Id:        4
Statistic   Train  Validation
    DOWN:  41.11%      48.65%
    NONE:  16.67%      13.51%
      UP:  42.22%      37.84%
   Count:     90          37
```

```
Node Id:        5
Statistic   Train  Validation
    DOWN:  33.33%      50.00%
    NONE:  30.77%      12.50%
      UP:  35.90%      37.50%
   Count:     39          16
```

```
Node Id:        6
Statistic   Train  Validation
    DOWN:  47.06%      47.62%
    NONE:   5.88%      14.29%
      UP:  47.06%      38.10%
   Count:     51          21
```

Figure 5.15: Decision tree using Probability Chi Square

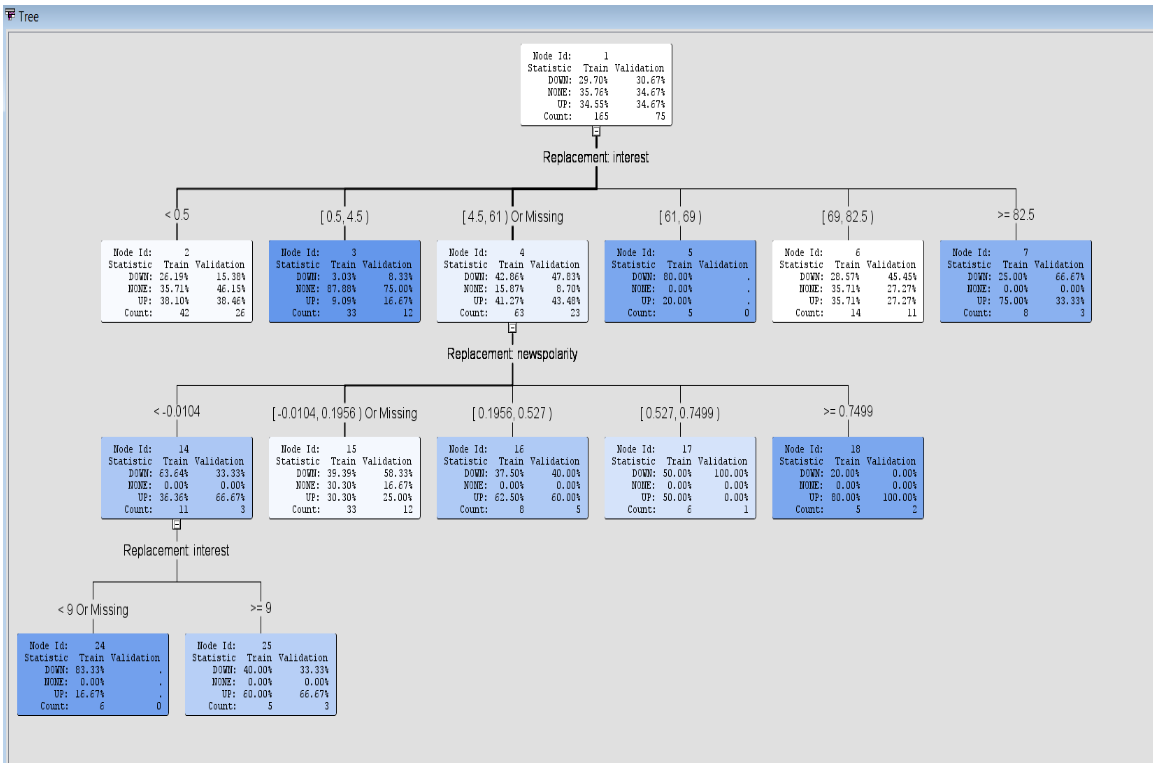
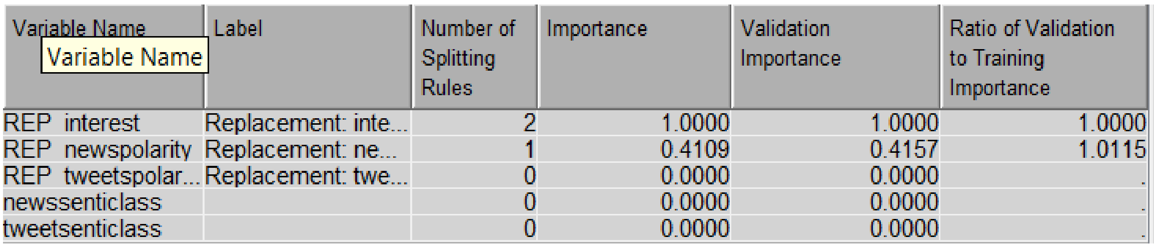

Tree
Node Id: 1 Statistic Train Validation DOWN: 29.70% 30.67% NONE: 35.76% 34.67% UP: 34.55% 34.67% Count: 165 75
Replacement: interest
< 0.5
[0.5, 4.5)
[4.5, 61) Or Missing
[61, 69)
[69, 82.5)
>= 82.5
Node Id: 2 Statistic Train Validation DOWN: 26.19% 15.38% NONE: 35.71% 46.15% UP: 38.10% 38.46% Count: 42 26
Node Id: 3 Statistic Train Validation DOWN: 3.03% 8.33% NONE: 87.88% 75.00% UP: 9.09% 16.67% Count: 33 12
Node Id: 4 Statistic Train Validation DOWN: 42.86% 47.83% NONE: 15.87% 8.70% UP: 41.27% 43.48% Count: 63 23
Node Id: 5 Statistic Train Validation DOWN: 80.00% . NONE: 0.00% . UP: 20.00% . Count: 5 0
Node Id: 6 Statistic Train Validation DOWN: 28.57% 45.45% NONE: 35.71% 27.27% UP: 35.71% 27.27% Count: 14 11
Node Id: 7 Statistic Train Validation DOWN: 25.00% 66.67% NONE: 0.00% 0.00% UP: 75.00% 33.33% Count: 8 3
Replacement: newspolarity
< -0.0104
[-0.0104, 0.1956) Or Missing
[0.1956, 0.527)
[0.527, 0.7499)
>= 0.7499
Node Id: 14 Statistic Train Validation DOWN: 63.64% 33.33% NONE: 0.00% 0.00% UP: 36.36% 66.67% Count: 11 3
Node Id: 15 Statistic Train Validation DOWN: 39.39% 58.33% NONE: 30.30% 16.67% UP: 30.30% 25.00% Count: 33 12
Node Id: 16 Statistic Train Validation DOWN: 37.50% 40.00% NONE: 0.00% 0.00% UP: 62.50% 60.00% Count: 8 5
Node Id: 17 Statistic Train Validation DOWN: 50.00% 100.00% NONE: 0.00% 0.00% UP: 50.00% 0.00% Count: 6 1
Node Id: 18 Statistic Train Validation DOWN: 20.00% 0.00% NONE: 0.00% 0.00% UP: 80.00% 100.00% Count: 5 2
Replacement: interest
< 9 Or Missing
>= 9
Node Id: 24 Statistic Train Validation DOWN: 83.33% . NONE: 0.00% . UP: 16.67% . Count: 6 0
Node Id: 25 Statistic Train Validation DOWN: 40.00% 33.33% NONE: 0.00% 0.00% UP: 60.00% 66.67% Count: 5 3

Figure 5.16: Decision tree using Entropy



Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
REP_interest	Replacement: inte...	2	1.0000	1.0000	1.0000
REP_newspolarity	Replacement: ne...	1	0.4109	0.4157	1.0115
REP_tweetspolar...	Replacement: twe...	0	0.0000	0.0000	.
newssenticlass		0	0.0000	0.0000	.
tweetsenticlass		0	0.0000	0.0000	.

Figure 5.17: Variable importance for Entropy based decision tree

Table 5.2: Event classification output

Set	Decision Tree 1		Decision Tree 2	
	True positive	True negative	True positive	True negative
Training	37%	67.8%	50%	76.2%
Validation	38%	69.7%	46.2%	77.8%

6 Conclusion

In summary, key takeaways from this project are diversified portfolio that comprise stocks of different risk categories is a plausible option. This can be identified through unsupervised clustering method, leveraging on the rate of return and volatility information of stock prices. Secondly, overvalued and undervalued stocks should be filtered out to focus investment on potential good valued stocks. This can be achieved by comparing P/E ratio against the company’s growth. Thirdly, news sentiment and interests (i.e. Google Trends search) is influential towards stock’s returns and should be monitored throughout investment period. Twitter sentiment need to be explored and enhanced further beyond this project, to properly study its effects towards stock prices. Decision tree that uses Entropy measure can predict the movement of stocks based on news sentiment and interests. A worth noting point is that the decreases in stock returns can be better predicted based on market sentiment and interests due to the concept of investor fear of losing money.

7 References

Allen, D. E., McAleer, M., & Singh, A. K. (2019). Daily market news sentiment and stock prices. *Applied Economics, 51*(30), 3212-3235. Retrieved from https://doi.org/10.1080/00036846.2018.1564115. doi:10.1080/00036846.2018.1564115

Andy, S. (2018, 21 September 2018). Five Powerful Ways To See If A Stock Is Overvalued. *Forbes*. Retrieved from https://www.forbes.com/sites/andyswan/2018/09/21/five-powerful-ways-to-see-if-a-stock-is-overvalued/#69476c8c6abe

Dash, S. R., & Maitra, D. (2018). Does sentiment matter for stock returns? Evidence from Indian stock market using wavelet approach. *Finance Research Letters, 26*, 32-39. Retrieved from http://www.sciencedirect.com/science/article/pii/S1544612317305111. doi:https://doi.org/10.1016/j.frl.2017.11.008

Eom, C., & Park, J. W. (2017). Effects of common factors on stock correlation networks and portfolio diversification. *International Review of Financial Analysis, 49*, 1-11. Retrieved from http://www.sciencedirect.com/science/article/pii/S1057521916301818. doi:https://doi.org/10.1016/j.irfa.2016.11.007

Wang, G. Y. (2010, 24-26 Aug. 2010). *Portfolio Diversification and Risk Reduction- Evidence from Taiwan Stock Mutual Funds.* Paper presented at the 2010 International Conference on Management and Service Science.