Risk Analysis and Default Prediction for Taiwan Companies

Milestone III Report

Team Member: Yuan-Hsi Lai, Pei-Ling Tsai

Current Status

In milestone II (The green bubbles in ***Figure1.***), we used sentiment analysis to give scores to posts retrieved from news and social media. These scores represent how confident the public is to the company. In this stage, we focus on forming a vector used for training. Thus, we combine the financial vector, sentiment factor in news and social media, relation factor (??) together. Then we used stock value with debt-to-equity ratio to form a training label for each company. After that, we used LSTM model, a kind of recurrent neural network, to train data.

A close up of a logo

Description automatically generated

Figure 1

**Accounting Data**

As mentioned in the previous report, we used nine financial ratios to form a vector. Then we used JSON format to save company name with its vector into a file (***Figure 2***.). Each JSON file means all the vectors in one quarter.

root@d5f90a8e4750:~/work/Documents/Big\_Data\_Class/project/fin\_data# ls ratios\*

ratios2018Q4.json ratios2019Q1.json ratios2019Q2.json ratios2019Q3.json ratios2019Q4.json

root@d5f90a8e4750:~/work/Documents/Big\_Data\_Class/project/fin\_data#

A picture containing black

Description automatically generated

Figure

**News and PTT Posts**

Modifications in the news crawler:

1. searching range: We changed 14 days of searching range to one quarter (3 months). To comply with the financial statements.
2. Adding filter: For some small companies that are not famous, the Google search engine would fill in unrelated articles to the search result, especially when the date is far from now. Thus, we used the filter to eliminate those articles. Although it would make the result list to be empty sometimes, it can prevent those unrelated articles from affecting the sentiment score.
3. Saving by quarter: To comply with the financial statements, we also put results into files by quarter.

Changed from Twitter to PTT:

Since only premium member of Twitter can get posts over 7 days, we decided to change the social media to PTT (a large social media in Taiwan).

About the PTT crawler:

(TBD)

A screen shot of a fence

Description automatically generated

Figure

A screen shot of a computer

Description automatically generated

Figure

**Relationship between Companies**

After building the graph database of company relationship, we transferred the “distance” between companies into a number. (TBD: What’s the usage of the number??) This number will join the final vector later.

**Improvement of Sentiment Analysis**

In the previous stage, we found that the accuracy of sentiment analysis didn’t do well. Thus, we improved by using the following methods:

1. …(TBD)

**Output labels**

We need labels to tell us how “good” the model is. Thus, we used some accountable information to help us. We put stock value and debt-to-equity ratio together, giving them equal weights. The reason why we used these two values is: the stock value is how investors think about this company and the debt-to-equity ratio is a straightforward number about how good the company can pay the money back.

**LSTM model**

Why do we choose this model?

LSTM model is a kind of recurrent neural network, which will “remember” the history information. Since we have data from different periods, the LSTM model will help us dispatch weights to different times, for example, it might give long-term data smaller weights.

Input shape

The LSTM model required 3-dimention input. Thus, the format is: (sample, timestep, factors)

* sample: each company
* timestep: each quarter (I set two timesteps, half year, in the model)
* factors: the final vector will all numbers we care

Activation

Currently, we use ReLu (Rectified Linear Unit) activation function to be our last layer, we will also try “Sigmoid” and “Tanh” to see which one best represents the data. (TBD)

Final Result

The current accuracy is not stable now. It may change from 32% to 90%. Thus, we are reviewing the configuration and algorithms, trying to reduce the total loss and improve the accuracy.

Challenges & Solutions

**Evaluating the Result (TBD)**

The rating report of companies is hard to get (need to pay a certain amount of money). Thus, we are going to track the trend of stock prices instead. The mindset is: If the stock price is rising in this year, that means investors are confident with the company and are willing to invest. This will generate more equity in the company and makes it less likely to default.

Next Steps

Training model (TBD)

References

\*1. Altman Z-score: <https://en.wikipedia.org/wiki/Altman_Z-score>

\*2. “Credit Rating Change Modeling Using News and Financial Ratios”: <https://www.researchgate.net/publication/262323524_Credit_Rating_Change_Modeling_Using_News_and_Financial_Ratios>

Appendix I. Coding Part