HappyAImen — MAFS6010U Final Project Report

LAM, Hiu Fung, 20476671

SHEN, Kairan, 20552932

WANG, Chenghui, 20552633

WU, Shukun, 20549545

XIAO, Yuxiang, 20551433

Department of Mathematics
The Hong Kong University of Science and Technology
Spring 2019

Abstract

This is a course project for MAFS6010U Artificial Intelligence in Finance at the Hong Kong University of Science and Technology. The project objective is to assist Alpha Intelligence Capital to track and monitor a group of private AI companies with the aid of Artificial Intelligence and Machine Learning techniques. This report documents the project progress, starting from ideas brainstorming to the final solution, as well as reflection on related academic articles about AI and finance. In the final deliverable, automation of sentiment analysis on the news of target companies would be conducted for monitoring purpose. Subsequent synthesis and suggestion would also be provided for further study.

1 Project Progress

During the project, sufficient communication is essential for the project team to understand and cater for the needs of Alpha Intelligence Capital. Two update meetings were held for progress catch-up and feedback from Alpha Intelligence Capital played an important role in determining the final solution.

1.1 Brainstorming

Brainstorming is always the most basic but important first step in the problem-solving process. At the beginning, thirteen approaches were came up to collect data and track progress of development of the target companies, including but not limited to different financial anaylsis on the companies and using Artificial Intelligence techniques on public available information.

With the feedback from Alpha Intelligence Capital, three approaches were selected for further investigation on the practicability of implementation. The first chosen idea was to apply voice-to-text transformation and natural language processing techniques on public production release conference. The intuition behind was that the production release conference always introduces the latest products and technology of the target companies, which in turn serves as an indicator for the company's techonology breakthrough.

The second chosen idea was to apply sentiment analysis on public and expert's response to evaluate the impact of the new policies on target companies. The intuition behind was that government policies usually have a tremendous impact on the industry prospects, especially in China.

The last chosen idea was to analyze the financial information of current clients of the target companies. The intuition behind was that an expectation on the companies' future profit could be formed, especially when the client relationships are mainly on project basis.

1.2 1st Update Meeting

Prior to the 1st update meeting, research on the three selected approaches were conducted. In order to facilitate efficiency, one of the companies, SenseTime, which is the 5th national AI platform in China, was picked as the target for trial investigation. Data example applicable to the chosen approaches were found, some as shown below briefly. SenseTime's public production release conference includes every year's "SenseTime AI Summit" and "World Artificial Intelligence Conference". Potential policies with impact on SenseTime includes a strategic alliance agreement signed with Shanghai Municipal Government. Current clients of Sense includes Oppo, Vivo and Sunning.

In the 1st Update Meeting, detailed research on the approaches were presented to Alpha Intelligence Capital. After exchange of views, the feasibility of the three selected approaches is questionable, mainly due to limited public available information on target companies and their clients. By common consensus, web scraping skills could be employed to obtain more news on the target companies automatically. Investigating into the deeper relationship between target companies and their clients, on project basis, could also be a direction for tracking purpose.

1.3 2nd Update Meeting

Prior to the 2nd update meeting, focus was mainly put on the suggestions came up from the 1st update meeting. Web scraping for news about SenseTime was done on PEdaily.cn, a renowned investment website in China. In addition, several SensTime projects, including Smart City project in West Bund of Shanghai and SenseRemote project with National Satellite Meteorological Center, were investigated. The research aimed to simulate the project tracking process by human effort, from the project initiation to the latest progress or final outcome of the project.

In the 2nd Update Meeting, the research result was discussed with Alpha Intelligence Capital. It was agreed that the inadequate transparency of project progress might be a huge obstacle for tracking and monitoring, especially when public information is not sufficient to determine whether a project is still in progress or already closed. Nevertheless, web scraping still serves as an useful tool for data collection, which could be the main tool in the project. Alpha Intelligence Capital also suggested that quantifying the economic bubble indicator, with the use of neural network, could be a direction, given that bubble is common among growing private companies.

1.4 Follow up

With respect to the idea of quantifying economic bubble indicator, the feasibility of implementation was assessed. Since economic bubble is an abstract concept, it could hardly be represented by a few factors. Assuming that bubble indicator rating system could be constrcuted, such factors might not be company-specific or accessible. As a result, this idea would not be considered as a final solution at the moment.

With an internal discussion, consensus was reached that a straightforward approach should be adopted. Making good use of the web scraping tool, news on the group of target companies could be obtained automatically. Therefore, data collection would not be an issue. Sentiment analysis could then be applied and the result of different companies could be compared and ranked. A better rank should represent that the public information indicates a positive view on the corresponding target company.

Sentiment Analysis, also known as Opinion Mining, is a field within Natural Language Processing (NLP) which constructs a framework to identify and extract opinions within text. Such system usually extracts attributes of the expression when identifying opinions. In solving a sentiment analysis problem, automatic methods, especially machine learning techniques, are used for most of the times. The sentiment analysis task is commonly modelled as a classification problem where a classifier is fed with a text and returns the corresponding category, such as positive, negative, or neutral. The modelling approach of setiment analysis is still evolving. Different sentiment analysis problems may be addressed by using different models. For example, a deep convolutional neural network, called Character to Sentence Convolutional Neural Network (CharSCNN), was proposed to exploit from character- to sentence-level information to perform sentiment analysis of short texts. More detailed information about CharSCNN will be discussed in Section 3.

2 Final Deliverable

In the final deliverable, five target companies were selected for anaylysis, including Sensetime, Malong, CloudWalk, Megvii and YITU. Articles on these five companies were crawled from 4 renowned news source website, Xinhua, Leifeng, PEdaily and Renmin. In total, 320 articles were collected for the subsequent text emotion analysis. The numbers of articles collected in different websites for different target companies are shown in Table 1.

Table 1: Number of articles collected on target companies in different websites

	SenseTime	**Malong**	**CloudWalk**	**Megvii**	**YITU**
Xinhua	29	10	29	15	8
Leifeng	39	7	25	35	5
PEdaily	18	2	3	10	9
Renmin	27	6	10	25	7

Sentiment analysis tool provided in the Tencent AI open platform was used in the project. At the current stage of project, due to time and data limit, Tencent's model, which is trained by powerful engines and millions of 100 billion social corpus, would be a more reliable tool. In the model output, a probability is shown on the positiveness on the input text, which is the emotion score. Effort was made on the automation on connecting to the Tencent AI sentiment analysis API and feeding all the articles for sentiment analysis. Emotion scores of all the articles were then obtained.

Sentiment Analysis Result

Based on the obtained emotion scores on each article, the average emotion scores were calculated for each target company on each news source. The result is shown in Table 2. The average emotion scores on these five companies is shown in Figure 1.

Table 2: Average Emotion Scores of target companies in different websites

	SenseTime	**Malong**	**CloudWalk**	**Megvii**	**YITU**
Xinhua	0.689835633	0.67566911	0.6692568	0.686085947	0.705076553
Leifeng	0.680941407	0.686683221	0.640162225	0.649663568	0.698378098
PEdaily	0.706337912	0.679484874	0.713388145	0.694513994	0.666450414
Renmin	0.679434227	0.689163993	0.646213084	0.677341714	0.669350607

As observed above, the average emotion scores of the five target companies are very close. SenseTime has the highest average emotion score. YITU comes the second and Malong comes the thrid. This implies that the media has relatively positive comments on these three companies. As such, a basic grouping of these five AI companies can be carried out based on the emotion scores. The first group includes Sensetime, Malong and YITU, the second group includes Megavii, and the third group includes CloudWalk. After analyzing the emotion scores on the companies dimension, the average emotion scores across different news source are investigated. The result is shown in Figure 2.

In Figure 2, the average emotion scores of SenseTime and Malong are more evenly distributed. It implies that the viewpoints on SenseTime and Malong are relatively consistent among different news sources. On the contrary, the average emotion scores of CloudWalk, Megavii and YITU deviate a lot in the four websites. It implies that the four news sources have different viewpoints on the companies. For example, CloudWalk has the highest average emotion score in PEdaily while it also has the lowest average emotion score in Leifeng. Such uncommon difference may indicate potential bias of

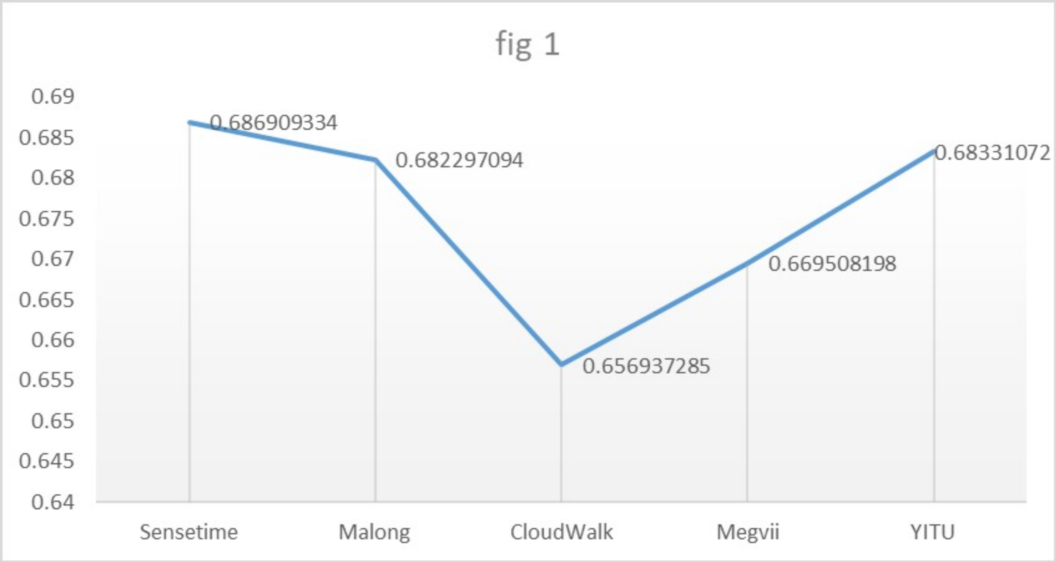

Figure 1: Average Emotion Scores of target companies

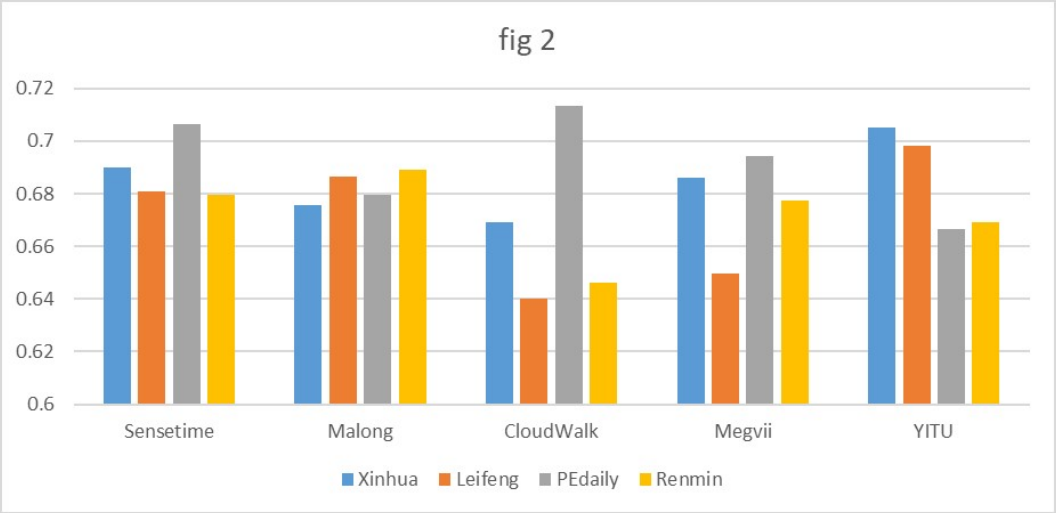

Figure 2: Comparison of the Emotion Scores

the news source towards CloudWalk, or exclusive news provided by a single news source. In both cases, this company should be monitored carefully.

To have a deeper look at individual companies, SenseTime is taken as an example and analyzed specifically. The number of articles on Sensetime is 114 in total. The distribution of the emotion scores on SenseTime is shown in Figure 3. The emotion scores evaluated on these 114 articles concentrate on the range of [0.6348, 0.7428]. When the emotion score of particular article falls out of this range, it may implies that special attention should be paid on that article which may contains extremely negative information about the company.

3 Reflection on academic papers

In this section, two academic articles on AI and Finance topics are reviewed. The articles selected are closely related to the project nature, sentiment analysis on news about private AI companies. Reflection on the articles could provide deeper understanding and inspiration on the project.

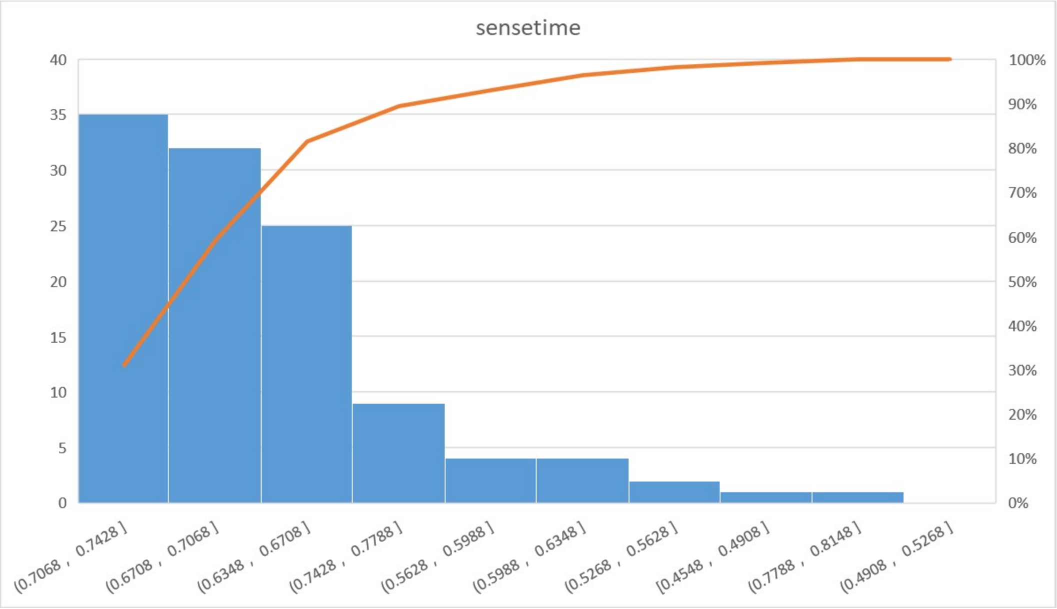

Figure 3: Distribution for SenseTime Emotion Score

3.1 AI article: Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts

Purpose To do sentiment analysis of short texts considering the limited contextual information that they normally contain.

Method A new deep convolutional neural network called Character to Sentence Convolutional Neural Network (CharSCNN) was proposed to exploit from character- to sentence-level information to perform sentiment analysis of short texts. The proposed network uses two convolutional layers to extract relevant features from words and sentences of any size, which technically means the word and character combined embedding as well as the sentence embedding. Given a sentence, the network with parameter computes a score for each sentiment label. In order to transform these scores into a conditional probability distribution of labels given the sentence and the set of network parameters, it applies a softmax operation over the scores of all tags, and finally makes use of stochastic gradient descent (SGD) to minimize the negative log-likelihood.

Data source CharSCNN was applied on two different corpora from two different domains: movie reviews and Twitter posts. The movie review dataset used was the recently proposed Stanford Sentiment Treebank (SSTb) (Socher et al., 2013b), which included fine grained sentiment labels for 215,154 phrases in the parse trees of 11,855 sentences. In experiments, it focused in sentiment prediction of complete sentences. However, it showed the impact of training with sentences and phrases instead of only sentences. The second labeled corpus was the Stanford Twitter Sentiment corpus (STS) introduced by (2009). The original training set contained 1.6 million tweets that were automatically labeled as positive/negative using emoticons as noisy labels. The test set was manually annotated by Go et al. (2009). In experiments, to speed up the training process it used only a sample of the training data consisting of 80K (5%) randomly selected tweets. It also constructed a development set by randomly selecting 16K tweets from Go et al.'s training set.

Result For the SSTb corpus, the approach achieved state-of-the-art results for single sentence sentiment prediction in both binary positive/negative classification, with 85.7% accuracy, and fine-grained classification, with 48.3% accuracy. For the STS corpus, the approach achieved a sentiment prediction accuracy of 86.4%.

Reflection The main innovation of this paper was that the word embedding in the input part was the splicing of ordinary "word embedding + character embedding". Character-level input was used to capture and combine information about affixes and roots. Character-level input is particularly effective in dealing with RNNs in sentence component tags, because '-ly' often means adverbs and '-ment' is common noun affix. This is conducive to the establishment of a more complete semantic model. In practice, weights can be set for the density functions of word and character respectively. Such combination of work-embedding and character-embedding provides improvement to the sentiment analysis performance. It could be a way to improve our project from the technical aspect.

3.2 Finance article: Venture capital and efficiency of portfolio companies

Purpose To investigate the inter-relationships between Venture Capital funding and portfolio firm performance.

Background Venture Capital funded companies show superior performance to non Venture Capital funded companies in general. However, since Venture Capitalists select and fund only the best companies, there are questions on Venture Capital's contribution to the performance of the companies they fund. As such, investigation was conducted on finding if the superior performance is contributed by the inherent characteristics of the firm or by Venture Capitalists after entering the firm.

Performance of VC funded and non VC funded companies The contribution of Venture Capital investment was shown in comparative studies of Venture Capital and non Venture Capital backed companies.The IPO and associated data was used in many studies to compare the performance of VC backed and non VC backed firms. It was found that VC backed firms have the below advantages:

- higher underwriter prestige, higher institutional holdings, and lower levels of under pricing for IPOs
- superior post IPO operating performance
- higher long term returns
- more innovative and associated with more valuable patents

In addition, a large fraction of the startups that made it to the public company stage were funded with Venture Capital, while very few companies received funding. By taking into account only true startup companies that go public, Kaplan and Lerner (2010) found that from 1999 to 2009, 60% of the IPOs had Venture Capital backing. Therefore, Venture Capital funding also significantly increases the success of a startup going public.

Reasons behind superior performance The two major activities of Venture Capital is screening and selection of investment opportunities and managerial inputs and value addition provided after investment. If the better performance Venture Capital funded companies is attributed to the former activity, it means that Venture Capital is able to pick the winner. In the other case, it implies the Venture Capital's ability to ensure the firm is managed well post investment.

A lot of studies conducted on the comparison of effect of selection and value addition were mentioned in the paper. Baum and Silverman (2004) analyzed biotechnology companies in Canada and found that characteristics that attract Venture Capital funding such as alliances, intellectual property are also associated with subsequent firm performance. However, human capital or top management characteristics of the firm that attract VC investment had little effect on subsequent firm performance. Therefore, this suggests a combination of both the selection and value addition roles in influencing portfolio company performance. Venture Capital is able to select companies that have strong technology and relationships, and provide management inputs that enhance the long survival of the firm.

In addition, Chemmanur et al. (2009) used Total Factor Productivity (TFP), which is the residual growth in output after accounting for changes in production factors as a measure to analyze the efficiency of portfolio firms. It is found that the efficiency of Venture Capital backed firms prior to receiving funding is higher than that of non Venture Capital backed firms. Further, the growth in efficiency after funding is greater for Venture Capital backed firms. This indicates the evidence for both the selection and the value addition effect. More interestingly, the contribution due to monitoring and value addition is found to account for a higher proportion of the increases in profits in Venture

Capital funded companies. 21% of the increase in profits are due to screening effects and 35% of the increase in profits are due to monitoring effect.

Reflection In the article, it is found that both the selection and value addition effect contributes to the superior performance of Venture Capital backed firms. Nevertheless, most studies indicate that the value addition effect dominates the selection effect. In view of this, a good tracking and monitoring procedure plays an important role for Venture Capital in determining the subsequent action on the portfolio companies. This also highlights the importance of the project topic.

4 Synthesis and Suggestion for further study

In the project, news about the five AI start-up companies on the authoritative website were collected through the web crawler. By conducting sentiment analysis on the news, the comparison of emotion scores gave some findings and inspiration for the monitoring purpose. Nevertheless, there are still rooms for improvement in the process. Here are some issues that may be addressed in further studies.

Collect more textual information Textual data is not necessary limited to authoritative news release websites. It is suggested that getting public opinions and comments from social media, such as Sina Weibo, Facebook, Twitter and Zhihu, can be one of the direction. While performing sentiment analysis on texts from different sources, weights can be added to the data from a more convicing source in order to get a objective result.

Time effect on the textual information It is not surprised that negative news about a company happen on occasion. However, when a series of negative news happens on a specific company or industry in a short time period, it may probably imply a warning signal. In view of this, further analysis can be conducted on the textual data at different time periods. Result can be compared horizontally to evaluate the public's opinion on the company during its development process or on the whole AI industry.

Advanced sentiment analysis model In the project, Tencent AI open platform sentiment analysis API was used for the sentiment analysis. However, it is suggested that a tailor-made sentiment analysis model can be employed in further study. In spite of the powerful engine of Tencent AI service, the sentiment analysis tool is designed for general use. A tailor-made model trained by comapny news may be more outstanding in performing the project task. Besides, advanced featured can be added to the sentiment analysis model. For example, CharSCNN introduced in the above AI article can be applied for sentiment anaylsis on public comments from social media.

Other than sentiment analysis In addition to sentiment analysis, some interesting ideas were came up during the work, two of which were further explored. The first idea is to build a bubble model by collecting company-related project progress, public relations articles, etc., to reflect whether the company has excessive or false reports at the current project stage. The second idea is to make use of the unsupervised deep learning approach to conduct clustering on the target companies. Data of listed companies in the AI sector can be used for model training. The target companies can then be clustered into different groups as if they are going to list in the future.

Individual Contribution

WANG, Chenghui Arrange the preliminary work; Assign tasks and cooperate with team members; Use Python to conduct web crawling to obtain relevant news of five target companies on four websites; Put forward the analysis ideas about the sentiment analysis result

SHEN, Kairan Use Python for data cleansing on news and articles about the five target companies obtained from four websites; Perform sentiment analysis with Tencent API; Analyze the sentiment analysis result and write the report based on it

LAM, Hiu Fung Research and analyze the selected ideas; Research and select AI and finance articles relevant to the project; Read the finance-related article and finish the reflection on it; Record and write the project progress in the report; Report documentation, including proofreading, organization and summarization, formatting, and all the necessary amendments and supplements

WU, Shukun Assist team members to collect company information; Assist team members on Tencent API connection in Python; Discuss with team members about the choice of news sources

XIAO, Yuxiang Find several news links of AI companies; Read the AI-related article and finish the reflection on it

References

[1] Cicero Nogueira dos Santos & Maira Gatti (2014) Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 69–78.

[2] Thillai Rajan (2010) Venture capital and efficiency of portfolio companies. In *IIMB Management Review (2010) 22*, pp. 186–197.