The University of Hong Kong

FINA4350 Textual Analysis and Natural Language Processing

Textual Analysis Applications in Predicting Credit Downgrades

Group 8

Chan Yin Tsung Lawrence (3035712169)

Chiu Tsz Chun Toby (3035712195)

Lu Yuzhen Steven (3035713943)

Li Hao Fiona (3035666308)

Wan Man Lau (3035666396)

1. Introduction

The prediction of credit ratings before they are publicly released by rating agencies is an immensely valuable tool. Credit ratings are a critical market indicator, revealing a company's financial health and influencing the broader economy through a domino effect. For example, the South Africa credit downgrade into junk bond status resulted in the currency dropping 11%, with long-term effects still being felt today. (Mukherjee, 2020). The most used method for predicting credit deterioration is fundamental analysis of individual companies, which uses conclusions drawn from financial data. While there are also machine learning solutions that predict credit deterioration, such as one used in a research paper forecasting the credit rating of decarbonized firms, such solutions lean heavily into using fundamental financial data (Yu et al., 2022).

This research aims to create a machine learning approach using natural language processing (NLP) to predict credit deterioration of companies based on publicly available information. The study focuses on the Chinese real estate industry for two reasons. The industry is unique for its high leverage, with a total of 117 billion USD worth of bonds set to mature in 2022, highlighting the market's size and importance (Reuters, 2022). Second, recent history has shown several instances of Chinese real estate companies and their corporate facing credit downgrades, providing ample examples for our training data. In this report, we will explore the use of NLP in predicting credit downgrades, thus contributing to the growing body of literature on the application of NLP in finance.

2. Methodology

This paper adopts a multi-step approach to building a model for predicting credit downgrades. Firstly, we will scrape the data using Selenium. Next, the data will undergo pre-processing before being used in two distinct machine learning models to draw insights.

Scraping

We selected downgrade reports from 30 different Chinese real estate companies with varying market capitalizations. We only incorporated downgrade reports from Moody's and Fitch as Standard & Poor's rating reports require a paid subscription. We recorded the rating decisions and the URLs of the websites containing the downgrade reports in a CSV file to prepare for web-scraping. The Moody's webpage requires a login every time it is loaded, and we added explicit waits that depended on triggers and interactive elements to the code to facilitate the login process and access the data. To scrape the reports from their respective webpages, we used Selenium to automate the login process since the requests package cannot handle JavaScript requests.

Preprocessing of Data

To clean up the information scraped from the webpage, we first kept only alpha-numeric characters. Second, we removed stop-words such as "a," "the," and "or", as well as one letter words to streamline the information further. Also, we conducted lemmatization to simplify the data collected and create more meaningful results in later steps. We removed specific words such as the credit rating company names, although some seemed to have slipped through. Finally, we utilized NLTK to tokenize the preprocessed words into our bag of words.

Model Building – Clustering Approach

The Clustering Approach involves two steps. First, we used the tf-idf model to identify significant words among the documents and convert them into a score or vector. Then, we inputted the results of the tf-idf bag of words to train the K-means clustering model.

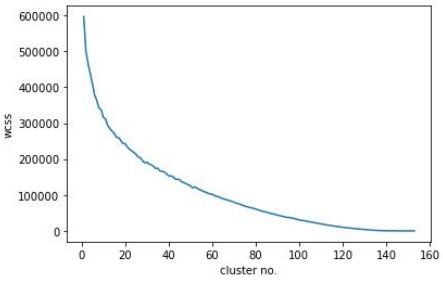
The tf-idf model is a statistical method used to evaluate the importance of a Bag of Words appearing within and across documents using the equation below:

$$tf - idf(term, document) = Tf(term, document) * Idf(document)$$

After obtaining a score for each term, we identified the most important and significant words among the documents using clustering. K-Means clustering is an unsupervised machine learning method that groups similar data points together without any manual input or guidance. This allows us to identify clusters of words that are important together and gain more insight

into which clusters of words can more effectively predict credit downgrades. To find the optimal k value, we used the elbow method, which involves plotting the average dispersion with k and identifying the point where the improvement in distortion declines the most. This method helps determine the number of clusters needed for accurate predictions.

Figure 1: Elbow method



After analyzing the results in Figure 1, it was challenging to determine the exact point where the improvement in dispersion declines the most. Instead, through careful visual inspection, we determined that the optimal number of clusters is k=5.

Next, to identify the best cluster of words that could predict default, we used our prior knowledge in credit markets to determine the most relevant cluster to use. This was a manual approach that relied on our understanding of financial markets.

Based on the significant words in the clusters, we apply them to testing documents to see if they are commonly used in reports and thus more likely to default. This approach allows us to gain deeper insights into the specific clusters of words that can better predict credit downgrades. However, the feasibility of this approach may be limited by the availability and quality of the testing documents.

Model Building – Similarity Approach

The Similarity Approach involves two steps. First, the Doc2vec algorithm is used to generate vectors from text data. This neural network-based approach generates vector representations for each unique word in a document by training over a large dataset. The vectors reflect the semantic meaning of the words based on their direction and magnitude. Words with similar meanings will have vectors pointing in similar directions.

Next, the soft cosine measure (SCM) is applied to calculate the cosine similarity of the vectorized word embeddings generated by Doc2vec. This similarity measure accounts for words that may not be identical but have similar semantic meanings. SCM can detect word similarity without words being mathematically close to each other in documents, representing an advantage over the regular cosine measure approach. The resulting similarity matrix clusters the words (represented as vectors through tf-idf weighted bag-of-words) with similar meanings together, enabling us to analyze which cluster of words more effectively predicts credit defaults.

Testing Data Collection

To evaluate the performance of our developed model, we acquired testing data from the same companies used in the model training. However, instead of using downgrade reports, we scraped the respective companies' annual financial reports. Specifically, we extracted only the Management Discussion and Analysis (MD&A) section, which contains a dense amount of text highly relevant to the firms' past performance and prospects.

To automate the extraction of MD&A sections from the PDF annual reports of selected companies, we implemented a script. The script utilizes the PyPDF2 package to extract text content from PDF files and applies a series of data cleaning and filtering processes to retrieve only the relevant MD&A sections.

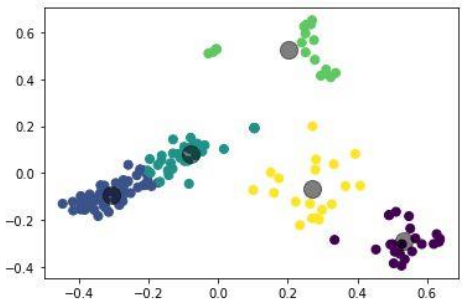
3. Results

Results from Clustering Approach

Figure 2: Word-Cloud Visualisation of Bag of Words

The results derived from the tf-idf bag-of-words model are shown in the word cloud, with words with a higher tf-idf score given larger font size. The 3 most prevalent words, "company", "moody", and "rating", do not provide any meaning for anyone looking for words that suggest deterioration in credit health. This was an error in the preprocessing, as these specific words should have been filtered out. There are promising signs however, such as the words "cash", or "liquidity", which could potentially be clustered together to generate a cluster of meaningful downgrade terms.

Figure 3: K-Means Clustering Visualisation



From visual inspection, we can observe 5 clusters of words, with center points of each cluster highlighted in larger grey circles. The clusters are quite distinct and far apart, representing an effective separation of groups of words in our training set. We then manually inspected each cluster to see which cluster is closely related to default, picking cluster 0 and 4.

Figure 4: Top words from each cluster

Cluster 0	*Cluster 1*	*Cluster 2*	*Cluster 3*	***Cluster 4***
company	advance	garden	fitch	**rating**
moody	rate	country	rating	**moody**
debt	applied	company	advance	**company**
liquidity	fitch	moody	cash	**entity**
next	value	rating	billion	**credit**
rating	rating	metric	fantasia	**road**
cash	inventory	next	rate	**rated**
operating	pud	margin	recovery	**king**
could	land	operating	value	**fitch**
billion	net	liquidity	land	**senior**

Results and discussion from Similarity Approach

After doc2vec, the documents have been converted into vectors.

Figure 5: Top words closest to default

Word	*Similarity Score*	*Word*	*Similarity Score*
outstanding	0.9575	experienced	0.9546
necessary	0.9476	forbearance	0.9189

fitch	0.9131	issuer	0.9128
solicitation	0.9046	consent	0.8976
dollar	0.8921	perpetual	0.8908
uncured	0.8853	removed	0.8842
journey	0.8806	avoid	0.8798
scenery	0.8788	commencement	0.8777
unsecured	0.8773	following	0.8767
senior	0.8662	derivation	0.8659

As seen above, some of the words such as “dollar” or “experienced” don’t give much meaning in the context of analyzing credit downgrades. This shows that even after obtaining results, we must manually filter irrelevant words from the dataset, in other words, removing noise. Regardless, there are words that are highly relevant to downgrades.

4. Overall Discussion

Figure 6: Final Results - Top 10 documents leading to credit downgrades sorted by Highest Similarity Score

Company Name	**Year**	**Cluster**	**Similarity**
Country Garden	2018	0	0.8018
Logan	2017	0	0.7788
Zhongliang	2021	4	0.7786
Shimao	2017	4	0.7778
Sunac	2019	0	0.7667
Zhongliang	2022	4	0.7557
Logan	2018	0	0.7522
Country Garden	2019	0	0.7395
Evergrande	2017	0	0.7310
Zhenro	2018	4	0.7242

Based on our assessment of 53 MD&A reports, our model can reasonably detect credit deterioration in company documents. All documents showed a degree of similarity between 50-80%, and mostly fell into clusters 0 and 4. The reports with the highest similarity scores within the correct clusters belong to a set of companies that suffered significant setbacks towards their credit situation during the Chinese Real Estate crisis.

Our results also show that both of our approaches have generated coherent results. The documents that show the highest similarity scores all originate from clusters 0 and 4, which matched our manual judgement of clusters. All generated similarity scores also fell within a reasonable band, which suggests our analysis was conducted properly and efficiently.

5. Implications and Application

The developed model has the potential to provide insights that human analysis cannot by pinpointing critical phrases and keywords that can reliably point towards credit downgrades. This is a promising step for NLP and machine learning applications in predicting credit downgrades.

One potential application of the model is credit valuation. Credit rating agencies are major market players, and their rating decisions can have profound impacts on the market and the wider economy. The additional insights that investors can gain from analyzing newly released financial reports give them an additional data point to aid in their decision-making process. Moreover, the speed of analysis from this model can allow investors to react quicker to report releases, resulting in more efficient asset pricing in the markets and improving the efficiency of the financial system.

Asset managers can also benefit from the application of this model. Since changes in credit rating affect the market price of the underlying securities, predicting credit downgrades can help expand the toolkit of investment managers. They can increase returns and minimize losses by divesting from securities in advance or using arbitrage strategies that front-run credit events.

6. Limitations and Further Research

Throughout the study, we have identified limitations in the technical aspects and the data we used to train our model.

Technical Limitations

Assumption of downgrades as a proxy for credit deterioration and potential default: Our training documents rely on the assumption that credit downgrades are the only event explaining credit deterioration. Our project should use more events such as credit default swap price spikes, and other corporate bond valuation measures to create a more comprehensive model.

Determining the best cluster in the cluster approach: we relied on our understanding of finance and credit to determine the best cluster of words that are close to default. To derive our results mathematically, we can use a logistic regression model to facilitate text classification. This will allow us to statistically separate clusters and find the optimal cluster that predicts default.

Efficiency: Many parts of the model are not fully optimized for speed and memory usage, which can result in scraping times of up to 40 minutes per run. Additionally, the machine learning model is not yet optimized for streaming, which puts a heavy load on RAM usage.

Scalability limitations: Our project is unoptimized for Big-O. As the data size increases, the time usage for scraping and memory usage during model building will increase exponentially at $\theta(n\text{^}2)$.

To improve on these limitations, the current code should be refactored to improve efficiency, which includes cutting out obsolete aspects of the code and incorporating multi-threading to allow for multiple scrapers to run at once. Secondly, the data pipeline could be improved to allow for streaming, namely integrating codes together and increasing compatibility between different parts of the code to allow for smoother data pipelines. Finally, more efficient data structures such as data frames could be utilized to provide faster access capabilities.

Data Limitations

Limited training data: Our training data is limited to credit downgrade reports by Moody's, Fitch, and S&P exclusively within the real estate industry in China. Although conclusive results can be drawn from this dataset, it is unlikely this model can be applied to other industries and geographies.

Effectiveness at analyzing other documents: it is unclear how effective the model is at analyzing other forms of public release (e.g., press releases, earnings reports) as they are significantly different from downgrade reports from rating agencies. This issue can be addressed by incorporating more training data into the model after resolving scalability issues in the technical aspect.

Further Research

Further research should build upon the current model in this paper, expanding the scope to more geographic locations/industries and incorporating different types of text into the training data to create a more robust model for predicting further credit deterioration.

Bibliography

Yu, B., Li, C., Mirza, N., & Umar, M. (2022). Forecasting credit ratings of decarbonized firms: Comparative assessment of machine learning models. Technological Forecasting and Social Change, 174, 121255.

Mukherjee, P. (2020, November 21). *'painful' downgrades will raise South Africa's borrowing costs, minister says*. Reuters. Retrieved May 4, 2023, from https://www.reuters.com/article/us-safrica-ratings/painful-downgrades-will-raise-south-africas-borrowing-costs-minister-says-idUSKBN2810FF

Reuters. (2022, January 19). *《图表新闻》中国房地产开发商在2022年将面临大批债务到期*. Reuters. Retrieved May 4, 2023, from https://www.reuters.com/article/china-property-debt-maturities-idCNL6S2TZ068