Application of Textual Feature Extraction to Corporate Bankruptcy Risk Assessment

by

Zhexuan Wang

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Systems Design Engneering

Waterloo, Ontario, Canada, 2017

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The inception of the Internet in the late twentieth century has established the ability to generate a huge volume of data from multitudinous sources in a very short period of time. However, most of this data is presented in an unstructured format. According to the latest research, unstructured data contains more comprehensive, effective and practical information when compared to structured data due to its descriptive characteristics, especially in finance, healthcare, manufacturing and other domains. It is anticipated that the effective use of data mining technology can be applied to the development of more accurate predictive models, decision-support platforms and man-machine interactive systems on unstructured data.

This thesis focuses on the application of a text mining system known as TP2K which stands for Text Pattern to Knowledge System, developed by my supervisor Professor Andrew K.C. Wong, to the finance industry. More specifically, the text mining system I proposed in this thesis is a concept-based textual feature extraction based on TP2K for corporate bankruptcy risk assessment. Bankruptcy risk assessment is to assess the bankruptcy risk of a corporation in the finance industry. It is linked to enterprise sustainability assessment, investment portfolio optimization and corporate management. Throughout the years, various models have been built using numerical and structured data (e.g. financial indicators and ratios). Yet no model has adequately leveraged the textual data for quantitative analysis in corporate bankruptcy risk assessment. Note that certain critical information such as strategic future directions and cooperate governance of an enterprise can only be reflected through textual data (e.g. annual financial reports). Recently, it has been reported that the combination of textual and numeric features will render a more accurate assessment of corporate bankruptcy.

Nevertheless, extracting features from textual data remains difficult since it still requires considerable human efforts. According to the existing literature, there is no obvious criteria for textual feature mining and extraction in finance due to the diversity of objectives and interests. From a general perspective, there is no simple criteria for textual feature mining and extraction in finance according to existing literature. Thus, domain experts still remain essential in the industry. The current textual feature extraction methods in finance can be categorized into two distinct types. The first type is based on a comprehensive handcrafted dictionary of proper keywords with continuous manual updating. The second type is based on data mining technology (e.g. high-frequency words). The former is time-consuming, while the latter usually produces results which are ambiguous, irrelevant or hard to be interpreted by industry in practice.

In this thesis, we (my supervisor and I) proposed a method known as concept-based textual feature extraction based on TP2K for corporate bankruptcy risk assessment. Compared to existing methods, this method can extract and mine textual features more accurately and succinctly from financial reports, allowing industrial interpretation in practice with limited human participation. It is semi-automatic and interactive. Its algorithmic procedure is briefly described as follows: (1) apply a linear-time and language-independent TP2K system to discover the Word, Term and Phrase (WTP) patterns from text data without relying on explicit prior knowledge or training; (2) apply a WTP-directed search algorithm in TP2K to find appropriate financial attribute names and their attribute values from the text context to obtain relevant attribute and attribute value pairs (AVPs) to build part of the Domain Knowledge Base (DKB) in support of predictive analysis of corporate bankruptcy risk. At the onset, domain experts will still play a major role in building the DKB. As more user-inputted domain information is integrated into the DKB, the system will become more automated to extract and validate related information for bankruptcy risk assessment with limited involvement from domain experts. In this thesis, AVPs have been used in corporate risk assessment to render more robust and less biased textual features. This allows experts to reasonably acquire and assist with the organization of individual selection rules in a comprehensive manner using traditional machine learning processing.

To validate the proposed method, experiments on financial data have been conducted. A collection of corporate annual reports containing textual and numeric information were adopted to evaluate the corporate risk assessment in a semi-automatic manner. Initially the extracted AVPs data was converted to binarized textual features in accordance with certain finance field criteria. It was then integrated with related numerical features (financial ratios) for traditional machine learning technologies to construct a predictive model for corporate bankruptcy assessment. The experimental results demonstrated an effective two-year ahead (T-2) prediction, outperforming prediction models based on only numeric features under 10-fold cross-validation. At the same time, we observed that all features discovered, numeric or textual, were consistent to the industry standard. Hence, we believe the proposed method has achieved an important milestone for assessing bankruptcy assessment in practice, and is potentially useful for providing trading advice for investors in the future.

Acknowledgements

I would like to express my gratitude to everyone who helped me throughout my Master's program.

Firstly, I would like to thank my supervisor, Professor Andrew K. C. Wong, and my co-supervisor, Professor Daniel Stashuk, for providing me with all the facilities for the research in past three years. For Professor Wong, I learned how to be a serious researcher and how to keep curiosity for the unknown fields in association with constant exploration. I also want to thank him on his constructive suggestions and his funding support in this process. I would also like to thank Professor Stashuk for the valuable advice in the past three years.

Secondly, I would like to thank my other two readers, Professor Fue-Sang Lien and Professor Shi Cao for spending time to be my reader and working on my thesis. Without their insightful suggestions, it would not be possible to further refine this research.

Thirdly, I would also like to thank my colleague Ph.D. Candidate, Antonio Szeto and Dr. Peiyuan Zhou for their selfless assistance and encouragement in my Master's program. They have given me invaluable insights and encouragement. At the same time, I would like to thank all of my department faculty members and staff for their support and course work. I am also grateful to my parents for their encouragement, support and attention in this venture.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ω-Text Well-Organized Text with line by line ordered sentences 15, 16

AVPs Attribute Name and Value Pairs 16

C-Pattern Hash Code-Based Patterns 16

C-Text Hash Code-Based W-Text 16

cTP2K the proposed Concept-Based Textual Feature Extraction based on TP2K 14, 15

DKB Local Domain Knowledge Base 5, 16, 20

HKEX Hong Kong Exchange and Clearing Limited 4, 25

TP2K Textual Patterns to Knowledge Software System 5, 15

W-Text Numerals and Common Words-Free ω-Text 15

W2C Table Word-to-HashCode Pairs Table 16

WTP Table Word, Term and Phrase Patterns Table transferred from C-Pattern 16

Chapter 1

Introduction

1.1 Problem Definition

In the finance industry, corporate bankruptcy risk assessment is widely utilized to forecast the possibility of enterprise bankruptcy [24]. This assessment is immensely important to individuals and institutions within the financial world because it offers an impression of control and stability.

To avoid unnecessary ambiguities, the term bankruptcy refers to the state of an individual, company or organization that is completely unable to resolve its statutory liability obligations. The bankruptcy risk is the measurement of the likelihood that an enterprise will reach bankruptcy as stated above [2, 3, 34, 41]. In general, the bankruptcy can be considered as a negative state of a company.

1.2 Background and Importance

There is a variety of types of risk that could lead to enterprise bankruptcy. They are briefly shown in Fig. 1.1 below.

The types of risk that are highlighted in blue can be measured through financial ratios. Correspondingly, the types of risk that are highlighted in yellow cannot be measured through traditional financial ratios in general. This means that they could only be evaluated through textual information.

Figure 1.1: Different Types of Risk leading to Bankruptcy

In practice, comprehensively reliable materials such as annual reports are commonly applied to evaluate certain types of risk which cannot be described through traditional financial ratios in corporate bankruptcy risk assessment. Here, annual report is the official comprehensive statement issued by the third-party audit concerning company activity from the preceding year. It is intended to give shareholders and interest groups comprehensive information about the operational conditions and financial performance of the company, especially its Director' s Report section and Management' s Discussion and Analysis section [7], where all of them contain a large amount of textual information.

Due to the global financial crisis in 2007, many financial institutions have suffered continuous loss in the past ten years. The impacted companies, in the tens of thousands worldwide, due to the continually worsening external environment and their weak internal risk management since that crisis, resulted in bankruptcy. Various sources speculate that this is just the beginning and such trends will repeat and intensify in the near future. According to information released in 2015 by The Companies Registry (CR) of the Government of Hong Kong Special Administrative Region (HKSAR), 121,963 registered companies were dissolved, bankrupt or written-off in Hong Kong. This was a 111% increase in comparison with 2014. For the groups of investors, lenders and company owners, this phenomenon is

a huge disaster.

Hence, bankruptcy risk assessment becomes a very significant and inevitable procedure to hedge against the latent risks in economic and financial affairs. Effective bankruptcy prediction is indeed very important. However, it is envisaged that a series of appropriate interdisciplinary theories, methods and approaches will be integrated together to assess or predict the upcoming future of enterprises; so as to accomplish the bankruptcy risk assessment comprehensively. Due to the complexity of the task, financial institutions normally use a large number of junior financial analysts to crudely analyze the basic foundation of the enterprises. It is a labor-intensive and time-consuming process.

Therefore, computerized intelligence systems are necessary to ensure efficiency and accuracy; to enhance the assessment or prediction and reduce time and labor cost in demand. Meanwhile, because of actual demand, bankruptcy prediction has been a hot spot among the academia, the industry and Financial Industry Regulatory Authority.

1.3 Limitation of the Existing Methods

Through multiple empirical studies, some achievements from the research of bankruptcy risk assessment have been generally validated and regarded as widely accepted routine process. However, current research on bankruptcy risk assessment is still restricted due to the following limitations.

First, in short, the corporate bankruptcy risk assessment system research is a typical binary classification problem in the field of machine learning when classes of assessment could be based on binary label assignment, i.e., bankruptcy and non-bankruptcy. Due to the complexity of the real market and economic environment, this binary prediction approach is unrealistic and is thus unable to meet demands.

Second, from the perspective of actual industry the financial indicators and intuitionistic symptoms of majority enterprises that are going to go bankrupt in the short term, e.g. within one year, are obviously severe. They can be investigated on the surface without a complicated assessment system. On the other hand, an enterprise may need an extended period from the onset of having financial problems to recognize bankruptcy as a series of problems which gradually manifest. Therefore, traditional bankruptcy and non-bankruptcy sample datasets are no longer appropriate.

Third, applied textual feature extraction methods in finance can be divided into two distinct types: methods based on a comprehensive handcrafted dictionary of proper keywords

with continuous manual updating, and methods based on data mining technology (statistic-based, e.g. frequency). The former type is time-consuming and has low efficiency [9], while the latter type usually produces results which are often ambiguous, irrelevant or hard to be interpreted by industry in actual practice.

Finally, to obtain bankruptcy risk assessment results that allow industry to interpret, it is necessary to involve domain experts and knowledge. However, due to the peculiarities of different datasets, the effectiveness of an assessment which solely relies on experts is undesirable, especially when large samples are involved.

1.4 Motivations

Bankruptcy risk assessment is a problem which is related to the financial market. Therefore, it should be interpreted from a finance perspective.

Researchers realize that there is a tremendous gap and intricate possibilities between bankruptcy and non-bankruptcy; hence bankruptcy risk assessment should be re-interpreted as a multiple classes (multi-level) classifications problem that is more adequate. In comparison with bankruptcy, examining the possibility that an enterprise may fall into bad financial conditions in the future has more practical significance.

Researchers also realize that concept-based textual feature extraction could bridge the gap between machine and human, and the output could be interpreted and analyzed further in a co-supportive manner. Effective bankruptcy risk assessment system should require machine intelligent auxiliary and supportive tools and their interface with domain experts and knowledge to overcome this hurdle. Yet today the analytic task of using machine intelligence is extremely difficult and requires considerable human effort.

1.5 Objectives

To bridge the listed gaps in this thesis, a multi-level bankruptcy risk assessment system is initially proposed to render more refined and detailed assessment results based on my research. There is a strong demand for a solution to this problem. In addition, blue chip companies, constituent stock companies of Composite Mid-Small Cap Index and listed companies with continuous losses in the past two fiscal years from Hong Kong Exchange and Clearing Limited (HKEX) have been used as samples in this research to replace the typical non-bankruptcy and bankruptcy binary classification problem in the previous research.

Meanwhile, this experiment is designed to investigate and demonstrate whether this newly proposed multi-level bankruptcy risk assessment system applies to real-world data acquired from the finance field and show how it can solve this problem and outperform the traditional binary prediction model.

Second, concept-based textual feature extraction for corporate bankruptcy risk assessment based on Textual Patterns to Knowledge Software System (TP2K) is proposed to discover concept-based textual features, i.e. critical feature names and their value from the unstructured context in some specific portions of corporate's annual report. Here, TP2K is developed by Dr. A.K.C. Wongs team and is in a patent preparation and filing stage. This research also attempts to include and implement the recommendations of domain expert recommendations at the onset rather than to rely solely on simple statistical methods (e.g. high-frequency words) to render more succinct interpreted meaningful textual features in the financial domain.

Third, the self-adaptive Local Domain Knowledge Base (DKB), is proposed as a component module in TP2K. It is briefly introduced and demonstrated to support the proposed multi-level bankruptcy risk assessment system. This is like an on-hand expert assistant because some procedures in this proposed system should be supported by expert advice or based on local domain knowledge. It also paves the way for future research on how to improve the service efficiency of expert domain knowledge.

The empirical results of this thesis can be utilized not only in bankruptcy risk assessment but also in enterprise sustainability assessment, investment portfolio optimization and corporate management suggestion. Financial institutions can make corrective decisions, avoid unnecessary losses, and reduce business operating costs at the same time.

1.6 Thesis Layout

The layout of this thesis is organized as follows. Chapter 2 reviews the relevant literature. Chapter 3 introduces the methodology and procedure of the proposed multi-level bankruptcy risk assessment system and concept-based textual feature extraction system based on TP2K. Chapter 4 presents experimental background, implementation, results and the analysis of bankruptcy risk assessment using proposed concept-based textual feature extraction based on TP2K. Chapter 5 concludes the thesis with suggested future work. A brief walk-through of the proposed concept-based textual feature extraction based on TP2K to corporate bankruptcy risk assessment is illustrated in Appendix A.

Chapter 2

Literature Survey

The research on bankruptcy risk assessment and prediction can be divided into three phases in accordance with its principal methodology: (1) the comparative study of traditional financial ratios between research objectives and industry standards in association with the utilization of intelligent empirical experience for judgment; (2) statistical modeling represented by univariate discriminant approach, multivariate discriminant approach and logit and probit regression methods; (3) computational tools and information technology based methods, e.g. machine learning techniques, expert systems, text mining techniques and genetic algorithm techniques. Many mature methods have been proposed and developed, particularly within academia. Some representative and momentous achievements are listed as below.

2.1 Representative Achievements

Firstly, most of the original datasets of bankruptcy prediction are essentially imbalanced data due to the relative small proportion of bankrupt enterprises in the real-world situation. In order to avoid the overwhelming lack of non-bankruptcy samples in the machine learning process, the number of bankruptcy samples and non-bankruptcy samples should be roughly equal when using pairing or other multi-step approaches. [51, 32]

Second, the feature selection for bankruptcy prediction is corroborated in bankruptcy risk assessment but has not rendered apparent positive improvements. In addition, features selection criteria are flexible and adaptable based on the characteristics of sample data. Since the dimensionality of bankruptcy datasets is relatively limited in comparison with

those in other fields and feature selection is usually time-consuming, this technology may not be required. [20, 69, 42]

Third, either machine learning or text mining techniques are sparsely related to the domain knowledge in the enterprise bankruptcy field. Although these are good calculation tools, they usually do not generate succinct explainable results. Therefore, experts with domain knowledge and experience must be involved in bankruptcy prediction research and assessment to make up this gap. At the same time, based on current research, the combination of expert knowledge and machine intelligence tools will improve the bankruptcy prediction accuracy. It will provide comprehensible results which the finance industry will accept. [35, 45, 64, 70]

Fourth, text mining technology which includes nature language processing, lexical, linguistic and semantic analysis and so on, can successfully extract textual features. Statistics-based textual features such as high-frequency words, which are difficult to procure and relate to industry practice from current qualitative information in financial data. These textual features can be regarded as the supplement of numerical features (financial ratios) from quantitative information. Numerical features are demonstrated as the foundation of modeling in general. Based on the current research achievements, bankruptcy prediction model will render more accurate classification results through combining numeric features (quantitative information) and textual features (qualitative information) together in comparison with those utilizing numerical or textual features independently. [8, 41, 31]

Fifth, common machine learning technologies, such as ANN, have been studied and utilized in the bankruptcy prediction field. It has been shown that different datasets often need specific optimal machine learning technologies to improve prediction accuracy. However, the performance of these optimal machine learning technologies are stochastic and dynamically changing and often unstable. They make current empirical studies that are not that persuasive. Nevertheless, we should mention here that the majority voting scheme has been used as a temporary solution to solve this machine learning problem. [61, 62, 47]

Sixth, according to the literature review, even for selecting appropriate features or using machine learning technologies for assessment, different sample data may need different optimal schemes. Indeed, there is no unified criterion for either textual-feature mining/ extraction or bankruptcy prediction modeling in the finance field. Most of the current bankruptcy prediction models are constructed based on a specific dataset, and cannot be applied universally with satisfactory results. Hence, it is unreasonable to compare assessment or prediction accuracy among different types or sets of sample data.

The mainstreams of the bankruptcy prediction research are principally focused on the following three areas: (1) bankruptcy prediction classifiers and algorithmic methods (2)

modeling featuring selection (3) textual feature extraction in empirical finance applications.

2.2 Bankruptcy prediction classifiers and algorithms

Empirical bankruptcy prediction classifiers and algorithms can be divided into two categories: statistical methods and artificial intelligence methods.

William Beaver [5]in his dissertation, "Financial Ratios as Predictors of Failure", initially put forward the univariate discriminant approach (UDA) to construct the assessment or prediction model using a single traditional financial ratio to forecast corporate financial status in rough. In 1968, Altman [2] first applied the Multivariate Discriminant Approach (MDA) in bankruptcy assessment and prediction research and proposed a general criterion to evaluate the probability of deterioration of corporate financial status, known as the Z value.

In fact, there are implications of many strict assumptions of MDA, such as linear separability and multivariate normality and the empirical studies which prove that most financial ratios cannot meet these requirements entirely. In order to overcome the inherent defects of MDA, the researchers introduced Logit and Probit Regression method in the later 1970s.

Ohlson [50]utilized the logit method in bankruptcy risk assessment or prediction problems with large database (2058 non-bankruptcy companies and 105 bankruptcy companies) and confirmed that the reasonable self-defined virtual features can assess the bankruptcy more accurately. This study was the first to use big database and non-financial ratios features as variables. It should also be noted that the binarized labeling method (giving 0 or 1 to replace actual numbers of selected variables) in association with subjectively pre-defined criteria applied on variables was initially proposed in this study.

Dimitras et al. [19]comprehensively reviewed all statistical techniques/ methods applied in bankruptcy risk assessment or prediction problems; and there was no widely accepted and verified new statistical technique satisfactorily applied in this field after that.

From the perspective of the traditional bankruptcy prediction model, Statistic Discriminant Analysis has always been a legitimate tool, mainly because of its interpretability and convenience. With the development of computer science and information technology, artificial intelligence methods are widely used in financial applications. Particularly, data mining technology is well recognized and supported by the industry. In comparison with statistical methods or classifiers, artificial intelligence methods are not constrained by any pre-defined assumption and can hence provide more accurate and faster prediction result when processing predictive analysis on complex databases.

Kumar et al. [36]collected and analyzed all undisguised literature since the 21st century in the financial field. According to the statistical distribution of their analysis, it is clear that support vector machine (SVM), naive bayes (NB), neural network (NN), decision trees (DT) and random forest (RF) are the first five commonly used and typically representative methods.

Except for the commonly used artificial intelligence classifiers or algorithms mentioned above, some other methods, like case-based reasoning [30], genetic algorithm [59, 33]and rough sets [46]were also used in the financial prediction or assessment research. However, relevant research of the above classifiers or algorithms has not been followed up by peers in the past 15 years and existing experimental results are lacking.

According to results obtained by Sun& Li [61]and Xiao et al. [67], the combination of multiple classifiers (at least two classifiers here) by majority voting can significantly improve the prediction results and reduce the error variance simultaneously in comparison with the single classifier method. The multiple class combination can further be subdivided into hybrid methods and ensemble methods.

Based on the above literature reviews and due to the differences between different benchmark databases in distinct experiments, we can bring out certain underlying intrinsic points. First, the same classifier may produce different individual optimal solution depending on the parameter setting for different benchmark database. Second, the prediction accuracy is not comparable among different database, or in other words, there is no practical significance. Finally, the research in finance prediction problems based on single classifier is sparse after 2012, and the combination of multi-classifiers by majority voting is more preferred nowadays.

2.3 Modeling feature selection

In the financial industry, domain experts have confirmed that some eccentric changes in enterprise' s numerical features (extracted from quantitative information, e.g. financial ratios) or textual features (extracted from any qualitative information, e.g. Directors Report) will be observed distinctly before the bankruptcy occurs veritably, which means it is rational to predict the bankruptcy through single or multi-classifiers method trained by these features.

The statistic-based features selection methods, such as Chi Squares, Info Gain and Filtered, cannot prominently improve the accuracy as reported in many current literature [17, 37]. They are time-consuming. Nevertheless, it is still adopted by many researchers. Moreover,

there is a perennial controversy between theorists and practitioners about which group of features is the most effective to train classifiers. Obviously, the optimal solutions could be different.

For numerical features selection, we find from various sources that applicable numerical modelling feature set can be procured by two major approaches based on: (1) universality; and (2) comprehensiveness.

For the first approach, Altman [2] proposed five numerical features which were the most discriminant variables and Pendharkar [53] adopted all five numerical features in Altmans work directly to ensure its practical significance in industrial perspective.

For the second approach, Liang et al. [41] and Zhou et al. [70] employed a total of 190 financial ratios and 208 financial ratios respectively extracted from the quantitative information as numerical features to construct the classifiers. These were the two highest number of numerical features used in a single experiment until the end of 2016. Although, the reported accuracy is not the best among those reported, the comprehensiveness of the final results was more persuasive than others.

For textual features selection of qualitative information from enterprise that included a series of contents relating to corporate governance, risk control, future prospect and so forth. This kind of contents cannot be quantified normally. Based on the consensus of present research, textual features set selection for bankruptcy risk assessment modelling are determined individually in accordance to the subjective preferences and intention of researchers.

Kim and Han [33] extracted six textual features from the qualitative decisions of domain experts including industry risk, operating risk, management risk, financial flexibility, credibility and competitiveness through genetic-algorithm-based data mining method, which has been demonstrated based on the practical experiences in South Korea principally. Sun and Li [63] summarized the latent qualitative risk factors within enterprises and assigned them into five features categories with experts experiences in China including market information, management and control, investment risk, consciousness and corporate governance; they even proposed a individual scoring criterion to evaluated its textual features selection. It is clear that, even textual features selection depends on experts experiences analogously in the above two cases, there is no universal accepted standard that is suitable for most situations to render a relatively optimal solution.

Alternatively, if researchers agree with certain textual features selected from the results of previous financial literature, they may use them directly and entirely and add or delete some subjectively.

The applicability of this scheme also has been evaluated by existing peer literature; for example, Cielen et al. [14] utilized part of features that were selected in Foster [22] and others works to structure their own features set.

Based on the current research, numerical features that are extracted through certain accounting standards are the foundation to construct and train classifiers legitimately in order to guarantee an understanding of the persuasive bankruptcy prediction results; while textual features are utilized as supplementary information to improve and enhance the bankruptcy prediction accuracy and comprehensiveness. This is achieved through the use and integration of the corporates' disclosed information in the market. Such attempts have widely demonstrated that combining numeric and textural features together would render more accurate classification results.

Due to different benchmark database, the similarity and comparability criteria are lacking in the existing literature. This leads to the strong subjectivity of optimal features group selection, even among numerical features, and inadequacy of their persuasion. On the whole, it is still an open question to determine which group of numerical/ textual features is the most appropriate and reliable in most of situations.

2.4 Textual feature extraction in empirical finance applications

From a general perspective, due to the difference in parties' objectives and interests, there is no simple criterion for textual feature mining/extraction in the finance field according to existing literature surveys. The current textual feature extraction methods, in finance, can be mainly divided into two distinct types: (1) methods based on a comprehensive handcrafted proper keywords with continuous manual updating; (2) methods based on statistic-based (e.g. high-frequency words) dictionary using data mining or text mining technologies.

For the method of the first type, the pioneering and the latest typical researches are listed below.

Lee and Yeh [39] employed 10 corporate governance indicators (CGIs) to predict financial distress. These CGIs were extracted from the board structure and ownership sections in Taiwanese enterprises' annual reports through referring to previous literature achievements or subjective judgments.

Shirata et al. [60] utilized certain parts of the audited annual reports of 180 listed companies (90 bankruptcy and 90 non-bankruptcy) in the Tokyo Stock Exchange from 1999 to 2005.

They successfully proposed a model to identify the most often appeared co-occurrences words and phrases associating with some professional key words that predefined by domain experts in accounting as textual features to the bankruptcy and non-bankruptcy of a corporation through CART-based SAF (a Japanese bankruptcy assessment system) model, and the similar method was also adopted by Hirokawa et al. [29].

Hajek and Olej [27] subjectively separated the words in enterprise' s annual report into six categories in accordance with its language functions through simplified sentiment analysis. These categories include: tenacity, accomplishments, familiarity, present concern, exclusion and denial. After some machine-learning-based validation examinations, the effect of each of the above category for bankruptcy prediction were confirmed. Those words from the most effective category were employed to construct textual features.

For the second method, some typical achievements in the main research direction in the past ten years are listed as follows.

Cecchini et al. [8] established a local dictionary named as Management Discussion and Analysis (MD&A)" to render the foundation of their bankruptcy prediction model through the combined method of statistics-based textual feature extraction (e.g. high-frequency words) and natural language processing (e.g. syntactic analysis and sentiment analysis). They proposed a model to predict two groups of financial events, bankruptcy and fraud, with a combination of textual and numeric features and rendered assessment results which were approximately more accurate than 83%.

Lin et al. [44] implemented a regression model to select the 6 most effective non-financial features, equivalent to textual features, out of 42 individually defined features from the original non-financial features dictionary. Therefore, features such as Shareholding of Board Members-Current vs. Prior Year, Ratio of Pledged Shares of Board Members, Shareholding of Board Members, Necessary Controlling Holding Shares, Other Investment Assets and Board Member Bonus to Pretax Income; these features were employed in their proposed financial distress prediction model.

Nugent and Leidner [49] proposed a supervised machine learning approach in association weakly-updated risk taxonomy and dependency tree analysis to conduct company-risk identification, which can provide a qualitative-based risk assessment as the output. They also attempted to clarify the relationships existing between different types of risks.

Based on the analysis of the existing literature about textual feature extraction methods in empirical finance applications, some key points are summarized as follows: (1) a domain dictionary, knowledge base or ontology are needed to support textual feature extraction, which are expected to extend automatically in the future research. (2) domain experts and their experiences are needed definitely for specific purposes in both experimentation and

research to evaluate extracted textual features; (3) there is a gap between statistics-based textual feature extraction or Natural Language Processing-based textual feature extraction and the practical application with industry interpretation; (4) most current studies focus on improving accuracy in academia which limits the practicality of this kind of research.

Chapter 3

Research Methodology

3.1 Overview of Principal Experiment Procedure

Taking into account the practicability, the procedure in this research is designed through two segments to demonstrate the proposed multi-level bankruptcy risk assessment based on concept-based textual feature extraction, which involve a feature extraction phase and a machine learning phase. The proposed concept-based textual feature extraction is based on TP2K which is designed to discover and extract textual features in the Feature Extraction Phase of this experiment. TP2K is referred to as the Textual Patterns to Knowledge Software System, an integrated software system developed by the research team of Professor Andrew K.C. Wong who is my Master Dissertation Supervisor. I am also one of the participants of the Invention. Since TP2K is generic and the methodology developed for my thesis can also be extended to solve similar problems, I would like to refer to the system proposed in my thesis as TP2K with a lower-case c in front like cTP2K. From hereafter, cTP2K will be used as equivalent to the concept-based textual feature extraction system based on TP2K used in my thesis above.

The experimental procedure to use the proposed Concept-Based Textual Feature Extraction based on TP2K (cTP2K) for concept-based feature extraction for corporate bankruptcy assessment will be briefly shown in the Fig. 3.1. The experimental procedure consists of two phases: Feature Extraction Phase and Machine Learning Phase.

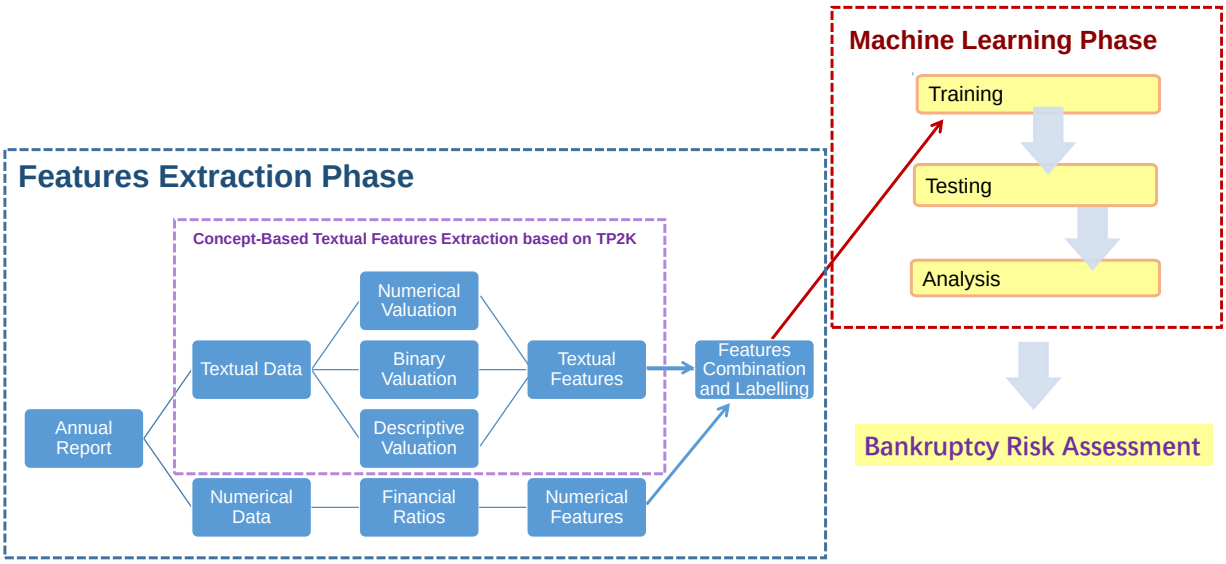


Figure 3.1: The principal experimental phases

3.2 Concept-Based Textual Feature Extraction based on TP2K (cTP2K)

An Overview of the proposed semi-automatic Concept-Based Textual Feature Extraction based on TP2K (cTP2K) is highlighted in Fig. 3.2 as follow.

In this thesis, cTP2K has been proposed and applied for corporate bankruptcy risk assessment. Here, according to [66], TP2K is an effective linearly time and language independent text sequence pattern discovery system. TP2K can discover Word, Term and Phrase (WTP) patterns from discretionary text or tabulated data with text and numeral contents without reliance on explicit prior knowledge or initial training at the onset.

TP2K is solely developed and owned by Dr. Andrew K. C. Wong and his team and is in a patent preparation and filing stage through WatCO of the University of Waterloo for technology commercialization. Due to the patent protection, only a limited technical description of this empirical study has been authorized to disclose.

In this thesis, cTP2K is divided into five systems in accordance with its function as shown in Fig. 3.4. We also used the following nomenclatures in the following description of this system: TP2K, Well-Organized Text with line by line ordered sentences (ω-Text), Numerals

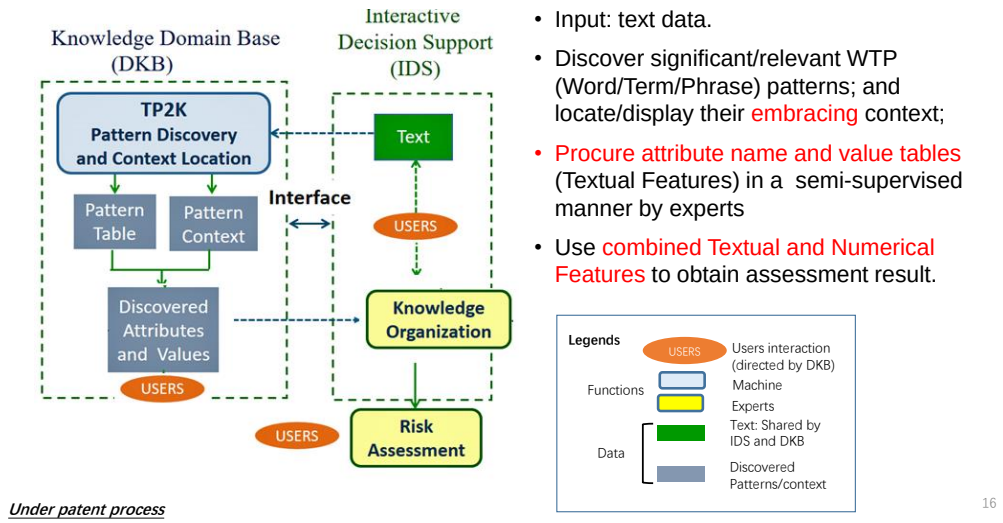


Figure 3.2: An Overview of the proposed semi-automatic Concept-Based Textual Feature Extraction based on TP2K (cTP2K)

and Common Words-Free ω-Text (W-Text), Hash Code-Based W-Text (C-Text), Word-to-HashCode Pairs Table (W2C Table), Hash Code-Based Patterns (C-Pattern), Word, Term and Phrase Patterns Table transferred from C-Pattern (WTP Table), Attribute Name and Value Pairs (AVPs), DKB, as summarized and presented in Fig. 3.3.

According to the step numbers presented in Fig. 3.4, the algorithmic procedure overview is briefly described as below.

3.2.1 Data Pre-Processing System

Step (1) The user can input any text document automatically. cTP2K arranges all sentences, line by line. We refer to this type of text file as ω-Text, which is converted from any existing word-document or extracted web-content on the Internet referred to just as Text. Hence, we have an automated process to translate the Text to ω-Text. Such step is referred to as: From Text to ω-Text.

Step (2) The cP2K will automatically eliminate numerals and common words [23] in ac-

Abbreviation	Description
TP2K	Textual Patterns to Knowledge Software System
ω-Text	Well-Organized Text with line by line ordered sentences
W-Text	Numerals and Common Word-Free ω-Text
C-Text	Hash Code-Based W-Text
W2C Table	Word-to-HashCode Pairs Table
C-Pattern	Hash Code-Based Patterns
WTP Table	Word, Term and Phrase Patterns Table transferred from C-Pattern
AVPs	Attribute Name and Value Pairs
DKB	Local Domain Knowledge Base

Figure 3.3: Nomenclature and Glossaries

cordance with a pre-defined set which is flexible and adaptable. The addresses of specific sentence in ω-Text and W-Text are the same. Then, the text only keeps the non-common words not in the default common words list. We denote such generated text without common words the W-Text. This step can be considered as a simple noise cleansing process while the original content contained in the ω-Text will be well preserved in the W-Text. This step is referred to as: From ω-Text to W-Text.

Step (3) A unique 6-digit hash code will be generated to each word entering the system through a special hashing scheme and documented in and extendable Word-to-Hash Code Table (W2C Table). Since some words are pertaining to a special group of W-Texts of interest and some are universal to the system, we introduce two W2C Tables :1) specific W2C Table, which consists of Word-to-Code pairs (W2C pairs) with frequency counts for specific domain (Finance); 2) universal W2C Table, which consists of Word-to-Code pairs (W2C pairs) with frequency counts for all words entering into the system for all W-Texts. The frequency counts of words in these two tables can be used for estimating the information relevancy measure later when necessary. Here, in essence, the relevancy of a WTP to an ensemble of ω-Text or W-Text in a study can be considered as information, universal rare and locally frequent. Then, every uncommon word in W-Text will be converted into 6-digit hash code to form a 6-digit hash code sequence with a spacing between each 6-digit hash code. We refer to that text as C-Text. This step is referred to as: From W-Text to C-Text in association with W2C Table.

3.2.2 Patterns Discovery System

Step (4) A Sequence Pattern Discovery Algorithm [66] is applied for discovering sequence patterns from the C-Text. These are code-based patterns, denoted as C-Patterns. These

Number of Step	System Module	Function Description
1	Data Pre-Processing System	Transfer Original Text into ω-Text
2		Transfer ω-Text into W-Text
3		Transfer W-Text into C-Text
4	Patterns Discovery System	From C-Text to C-Pattern
5		From C-Pattern to WTP Table (Word-Tern-Phrase)
6	Pattern Embedded Context Finding & Locating System	Find & Locate Context through WTP Table
7		Find & Locate Context through Domain Knowledge
8	Attribute Name and Value Pairs (AVPs) Extraction System	Validate Attribute Name
9		Validate Attribute Value
10		Attribute Name and Value Pairs (AVPs) Table
11	Self-Extended Domain Knowledge Base (DKB) System	Construct Local Domain Knowledge Base (DKB)

Figure 3.4: Detailed descriptions of systems in cTP2K

discovered patterns are made up of digit units corresponding to the W2C Tables. This step is referred to as: From C-Text to C- Pattern Table.

Step (5) All digit units with length less than 6-digit of the code-based patterns, especially at the beginning and the end of a potential pattern, will be abandoned since they are not parts of word patterns. After removal, all the retaining digit units of code-based patterns containing a sequence of Hash Code will be converted back to words corresponding to universal W2C Table. Then a Table of significant Code-patterns and so are W-patterns that are made up of significant Word, Term and Phrase patterns will be discovered. They will be included in the extending Word, Term and Phrase Table abbreviated by WTP Table. In the WTP Table, each pattern and its frequency of occurrences, statistical significance, domain relevancy, digit length and the sentence addresses in the W-Text containing it will be registered. This step is referred to as: From C- Pattern Table to WTP Table.

3.2.3 Pattern Embedded Context Finding & Locating System

All patterns in WTP Table are essentially extracted from W-Text, which can be considered as noise free -Text. As mentioned in Step (2), the ω-Text containing the original content of W-Text is well preserved.

Step (6) The sentences embracing any pattern in WTP Table of W-Text can be found and located in the original ω-Text without any information loss due to the sentence addresses (sentence numbers) correspondence between W-Text and ω-Text. We abbreviate such process by WTP-directed search. Furthermore, as the derivatives of WTP-directed search, co-occurring WTPs can also be retrieved from the specific context in the ω-Text by setting a precondition. This step is referred to as: From WTP Table to ω-Text.

Step (7) In order to ensure the acceptability and interpretability by industry, Pattern Embedded Context Finding & Locating System embedding any WTP (such as financial terms, concepts or criteria) from WTP Table or Domain Knowledge Base initiated by the users and extended by cTP2K could be used to direct the search and locate the WTPs and all their contexts which contain the WTP. Hence, any significant domain specific word/term/phrase or a furnished list of words/terms/phrases can be individually predefined by the users or domain experts as searching concept or criteria. The financial criteria directed search is a complementary function to overcome conceivable patterns which could be misinterpreted in WTP Table.

Either WTP-directed search or financial criteria directed search can simultaneously locate and display the entire relevant context in the ω-Text for interpretation, confirmation and assurance before the concept enters into the Domain Knowledge Base. Thus, this is a semi-supervised method to allow the domain experts to have candid assurance and the final say.

3.2.4 Attribute Name and Value Pairs (AVPs) Extraction System

Step (8) All WTPs and individually predefined financial criteria are available to render appropriate financial attribute names that should be further explored and selected by the user or domain experts in association with their task objectives. In this empirical study, a predefined concept-based local list of key attribute names extracted from subtitles of Director' s Report or Management Discussion and Analysis is provided. Users can upload the existing list directly or establish their own attribute names list. The Attribute Name and Value Pairs (AVPs) Extraction System will automatically update the local list when new inputted attribute names are added into it in compliance with the users' personal preferences. Such step is referred to as: Attribute Names Validation.

Step (9) Once attribute names are determined, the corresponding attribute value can be determined distinctly from the located contexts that contain the attribute name. When an attribute name and its embedding contexts are displayed, its relevant attribute value (AV)

could be scrutinized by the domain experts or fostered by the Domain Knowledge Base to enter into or refute from the Attribute Name and Value Pairs (AVPs) Table for subsequent pattern analysis. To ensure that only relevant information will get into the system, the users or the domain experts are required to select qualified values to pair with attribute names from context through subjective corporation expectation. Such step is referred to as: Attribute Values Validation.

Step (10) Combining attribute names and attribute values from the text context can obtain relevant attribute name and value pairs (AVPs). An AVPs Table will then be obtained for each separate inputted ω-Text. These AVPs are individually selected textual features that will be combined with other numerical features in the subsequent machine learning phase. Such step is referred to as: From AVPs to Textual Features.

3.2.5 Self-Extended Domain Knowledge Base (DKB) System

In this thesis, the AVPs have been used principally in corporate risk assessment. They will render more robust and less biased textual features for experts to reasonably acquire and organize into individual selection rules in a comprehensive manner through machine learning. Therefore, acquired AVPs tables are only a small part of the Domain Knowledge Base (DKB) in support of predictive analysis of corporate bankruptcy risk. Furthermore, some predefined and self-extended sets/ lists/ dictionaries, e.g. attribute names list, are also utilized in the concept-based textual feature extraction based on TP2K (cTP2K) system. All these sets/ lists/ dictionaries can also be considered as a part of the extending DKB.

In the beginning, the user or domain experts will play a major role in building the DKB. As more discovered or user inputted domain knowledge is acquired and integrated into the system, the system is supposed to become more automated and independent to extract and validate related additional information for bankruptcy predictive analysis with less and limited domain experts' involvement.

According to the cTP2K design, the current Domain Knowledge Base (DKB) System has been developed for man-machine interface. In the future research, the acquisition of AVPs could be organized into relational datasets for Attribute Value Association (AVA) Discovery or Mixed-Mode Pattern Discovery (both developed by our team) for multi-objectives analysis from the unstructured data (textual and tabulated data).

An entire walk-through is added in Appendix A at the end of this thesis to illustrate how this proposed Concept-Based Textual Feature Extraction based on TP2K (cTP2K) works on a real sample in this experiment.

3.3 The Use of TP2K for Concept-Based Textual Feature Extraction

There are two significant points that need to be further emphasized here.

First, due to the unambiguous specific objects of bankruptcy risk assessment, cTP2K is solely used to discover and extract concept-based textual features, i.e. critical feature names and their value, from the unstructured context, and to disclose them in a succinct and comprehensive way to the users for in-depth corporate bankruptcy risk assessment. However, in practice, TP2K is a comprehensive system with multiple functions, such as supportive knowledge organization and its ability to produce pattern analytical results from Mixed-Mode Relational Data derived from the unstructured text.

Second, in this empirical study, the usefulness of cTP2K for corporate bankruptcy risk assessment has been demonstrated. From the generic nature of the problem, this concept-based textual feature extraction method is not restricted to the finance field since there are also a large number of textual data with similar nature existing in many other fields such as healthcare, energy, ergonomics, manufacturing, and others. They too can use a concept-based textual feature discovery and analytic system such as TP2K.

Chapter 4

Experimental Background, Implementation, Results and Analysis of Bankruptcy Risk Assessment Using cTP2K

A diagrammatic description of the experiment process can be referred to as the diagram given in Fig.4.1 below.

4.1 Experiment Background and Implementation

4.1.1 Raw Data Collection

Data Source

In this empirical study, corporate annual reports have been utilized as data to support all the experiments; and all corporate annual reports are collected from Hong Kong Exchanges and Clearing (www.hkexnews.hk) directly, abbreviated by HKEX.

There are three reasons to explain the appropriateness of this data source.

First, HKEX assumes that all investors in this market are presumed to be rational investors, and there is no exact official risk warning mechanism based on the company's

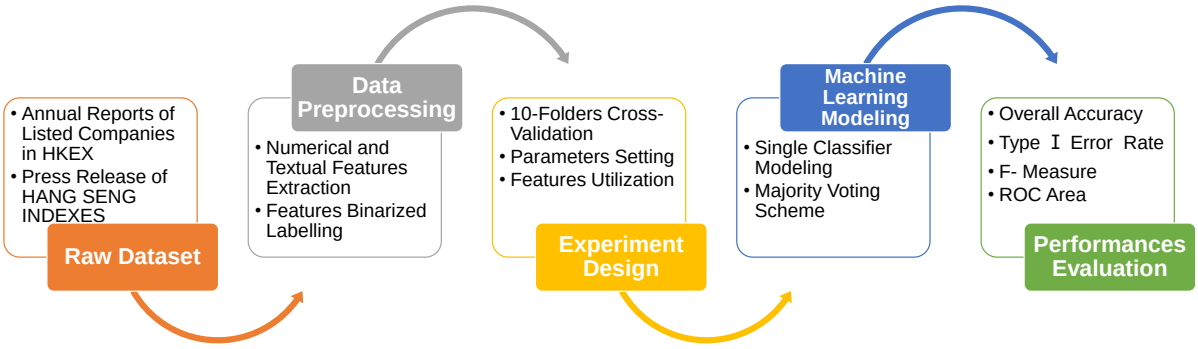


Figure 4.1: The principal process of the experiment

financial position and previous performance adopted or documented by HKEX. Therefore, the proposed bankruptcy risk assessment based on the data of HKEX has realistic instructive significance.

Second, listed companies, including main board and growth enterprise market (GEM), must overtly display their annual reports in accordance with the relevant provisions in the official website of HKEX within the prescribed time point; as a result, the data is more reliable and convincible.

Third, HKEX, is a well-known and fully-fledged financial market, and has been completely investigated by investors from all over the world through different industry criteria. These industry judgments will assist the users to verify the correctness of the results generated in this empirical study and will further improve the practicability of the proposed bankruptcy risk assessment method in the future.

Forecasting Time Span

The forecasting time spans of raw data (annual reports) in this empirical study is two years ahead, which is the most generally accepted time span of bankruptcy risk assessment in current research. It implies that the annual report in the year T-2 will be utilized to assess the company's financial performance in year T.

Different from the typical European and American financial markets, the fiscal year of listed company in HKEX is not necessarily between January 1 and December 31 each year. At the same time, the annual report should be released within four months after the end of a fiscal year. Therefore, all annual reports are drawn from fiscal year 2014 or 2015 in association with its discernible financial performance in the year 2017.

Three points should be clarified about forecasting time span selection:

(1) The annual report is stipulated to disclose the prospect and programme of the enterprise in the short-term future, which means that there is no theoretical support for more than three years medium and long-term assessment based on the annual report in this field. (2) There is a four-month lag at least between the end of fiscal year and annual report releasing, which means that one year ahead bankruptcy assessment is only eight months ahead at most realistically. It is not sufficient in the real-life practice [1]. (3) Referring to the widely recognized research achievements, like Z Score Model [2], two-year ahead bankruptcy assessment is the major object with theoretical support in this field.

Bankruptcy Risk Categories

In financial field, combining with advanced machine learning technologies, companies can be classified into three categories in accordance with their low, standard or high bankruptcy risk based on their financial ratios and other supportive materials theoretically [25, 52]. According to the possibilities of their bankruptcy, the data of listed companies used in this empirical study can be categorized into three groups which are briefly described as below.

For the first category, Blue Chips Companies are referred to the companies as those having obvious larger margin of safety in association with comprehensive positive fundamental factors. The Hang Seng index is calculated by the combination of the market values of the fifth selected representative enterprises, which are official Blue Chip Companies in Hong Kong. Due to their stable profitability, these companies are considered to be mostly risk-free as the convention in the industry.

For the second category, some constituent stock companies of Composite Mid-Small Cap Index, which have no loss in the fiscal year in association with the Return on Equity (ROE)

of them are lower than average rates of all listed companies in HKEX. This means that these companies failed to provide actual return for their shareholders in the last fiscal year and would negatively affect the investors' confidence. The bankruptcy risk of this kind of companies is considered as standard in this experiment.

The third category consists of listed companies having two or more than two years consecutive loss by the end of the available fiscal year or under the Delisting and Suspensions Processing by HKEX. These companies are suffering from enormous financial difficulties for a period of time, and their possibilities of bankruptcy is relatively higher than those of the other two categories.

The list of companies of the first two categories will be re-evaluated and released by Hang Seng Indexes quarterly, and yearly for the last category. In this experiment, these listed companies give the credential as sample enterprises for discernible financial performance evaluation.

Category Data Stratified Sampling

Here, the thing that should be noticed is that there are big differences between the numbers of companies in the above three categories in real practice. For example, the number of Blue Chip Companies is fixed to 50 and does not change for a long time; in the same period, there are more than 160 qualified sample companies in the third category by the end of fiscal year 2016. This phenomenon will definitely lead to the imbalance data problem and produce unnecessary deviations and bias in reducing the accuracy of the predictive results in machine learning.

Therefore, stratified sampling, initially proposed by Altman [2], was used to extract the same number of positive samples and negative samples in order to overcome this hurdle. As stratified sampling has been applied in this study, the numbers of companies in the above three categories are the same in our experiments.

Dataset Description

The dataset consists of the listed companies which satisfy all the above preconditions and have available data in the meantime.

The dataset consists of 114 companies in total. They are: (1) 38 Blue Chip Companies in HKEX; (2) 38 constituent stock companies of Composite Mid-Small Cap Index; (3) 38 listed companies with the consecutive two fiscal years loss before 2017. The discernible financial

performance of these 114 companies in year 2017 is referred and confirmed through the latest press release of HANG SENG INDEXES on May 19th, 2017. Here, due to different accounting standards applied on Bank, Bank is not included in the above 114 samples.

If we consider only the total of 1973 listed companies in HKEX by the end of 2016, it is notable that the applied dataset sample consists of nearly 6% of all the listed companies in HKEX.

4.1.2 Feature Selection

Numerical Feature Selection and Binarized Feature Processing

Referring to the literature on corporate bankruptcy assessment between 2012 and 2017, we find a total of 25 financial ratios initially selected and utilized as numerical features [16, 54, 4, 10, 13, 15]. They comply with the requirements of financial procedures by accepted accounting principles in Hong Kong. These numerical features can be broken down into five categories to comprehensively depict the financial conditions of the companies, namely, Liquidity-Solvency, Investment Value, Capital Structure, Profitability-Growth and Operation Capacity. Some ambiguities would be generated if only a certain financial ratio is used solely. Therefore, there are at least three complementary financial ratios in each category above.

In order to train the machine learning algorithms correctly, in general, there are two methods to assign class labels to the numerical features in finance research. They are the Min-Max Normalization or Binarization based on standard values of financial ratios.

For the former method, all values of numerical features will be normalized through mapping raw data into the interval [0, 1] through linear transformation, and its simple circulation as follows [41]:

$$Y_i = \frac{x_i - min_x}{max_x - min_x} \tag{4.1}$$

where x= (x_1..., x_n) is the raw dataset, and Y_i is the corresponding normalized data of x_i. However, the latent knowledge of this statistics-based Min-Max Normalization is almost irrelevant in connection with the finance field.

Hence, in view of this, the latter method, Binarization [50], has been adopted in this study.

Here, an example will be provided to clarify how this method works to assign class label to numerical features. Suppose that the current ratio is a liquidity indicator that measures the resources of a firm to meet its short-term obligations. Its formula is defined as:

current ratio = current assets/current liabilities. In accounting practice in Hong Kong, generally speaking, a current ratio less than 2 would indicate some underlying difficulties of a company to fulfill its short-term obligations. Based on such standard, in this study, if the number of current ratio is greater than 2, we consider it as positive and labeled it by 1; else as negative and labeled it as 0.

Here, there are certain points that should be emphasized in this study: 1) not all the financial ratios can be roughly estimated through standard values, especially for some elastic ratios which may not be adaptive for this binarized processing; 2) these recognized standard values are defined by the practitioner in the market adaptively, like the above current ratio, or fixed based on textbook, like Price-to-Sales Ratio [21]; 3) generally accepted accounting principles in Hong Kong particularly referred in this study; 4) different industries may have distinct recognized standard values on specific ratios, but this factor will be ignored to avoid unreasonable small data volume [29].

In order to ensure the correctness of the result, all numbers used to calculate the financial ratios in this study are extracted from Balance Sheet, Income Statement and Cash-Flow Sheet in Corporate Annual Report directly. A total of 25 financial ratios have been calculated as numerical features in this experiment, and 20 of them have been used finally. As shown in Fig. 4.2 below, five financial ratios marked in red were abandoned for the following reasons: 1) the relevant data is temporarily incomplete; 2) the ratio is not adaptive for some sample companies; 3) there are similar ratios, which can be utilized to measure the same characteristic of the financial status of enterprise.

All the detailed information about numerical features, including its financial ratios names, symbols, formulas, categories and binarized feature processing has been organized into a table which is provided in Appendix B at the end of this thesis.

Textual Feature Selection and Binarized Feature Processing

A total of 10 textual features, which are also subtitles in Director's Report or Management's Discussion and Analysis, have been attempted to extract in the specific portions of corporate annual reports through the proposed concept-based textual feature extraction based on cTP2K in this thesis. In addition to the financial performance, we would like to find out the corporate governance, market circumstances, expansion capacity, underlying risks, future prospect and so forth in this section. These can be considered as the minimum essential information of a corporate.

Due to the quantitative information that has been represented as numerical features through financial ratios, the first priority of textual feature selection is to assign qual-

itative information that cannot be expressed as numerical features. The main objective here is to use information in the Director's Report and Management's Discussion and Analysis adequately and comprehensively while avoiding information loss as much as possible. There is a premise, that is, all the released information in the corporate's annual report should be significant to evaluate the enterprise performance.

About binarized processing, referred to the current research, valuations of textual features can be divided into three types as below.

First, the textual features which have exact numerical value in the last two years, like the number of employee. Usually, the comparative status of the employee number in two consecutive years is adopted, i.e. if the number of employees is more than that of the last year, the employee label could be considered as positive and assumes label 1; otherwise, as negative with label 0.

Second, the textual features have essential analogous binary value (Yes/ No). For instance, for dividend, if it has been mentioned in the Director Report as yes, it could be considered as positive and the label will be given the value 1; otherwise, as negative with label 0.

Third, the textual features are description statements without any existing specified value, like future prospect, underlying risks and market circumstances. Hence, the users or experts are required to interpret some key contexts in association with some corresponding important attribute names from financial knowledge, and assign label 1 as positive and label 0 as negative in a subjective manner. As knowledge accumulates, a synonymous synthesis may be able to map them into a category confirmed by experts in the future research.

A total of 10 textual features have been attempted to collect in this experiment, and 6 of them have been finally used as shown in Fig. 4.3. For the abandoned textual feature T7 and T10 in Fig. 4.3, their evaluation criteria were not clearly related to current research or accounting domain knowledge. For the abandoned textual feature T8 and T9 in Fig. 4.3, it is not applicable for all sample companies in this research.

All the detailed information about the above textual features, including names, symbols, representative corresponding important attribute names, and binarized feature processing, has been combined with the numerical features and also organized into a table which is provided in Appendix C at the end of this thesis.

Output Results

As the study above explains, according to the possibilities of bankruptcy, blue chip companies are referred to as approximately risk-free in anticipation which would be labelled

2; constituent stock companies of Composite Mid-Small Cap Index in this experiment are considered to have standard risk and labelled 1; listed companies which have two or more than two years with consecutive loss or under the delisting and suspensions process by HKEX would have relatively high-risk and are labelled 0.

Feature Selection Methods

There are three principal objectives of Features Selection listed as follows: 1) increasing efficiency by reducing the data dimensionality; 2) removing redundancy; and 3) selecting the most discriminative features based on the specific database. However, the characteristics of financial data are their relatively small dimensionality in comparison with those in another field, e.g. bioinformatics, and the improvement of prediction through features selection has been proved obvious and effective up-to-date in corporate bankruptcy risk assessment. Therefore, the Feature Selection will not be investigated in this study as the main objective.

4.1.3 Predictive Classifiers

Alternative Classifiers

Naive Bayes (NB) [40, 26, 48] is inspired by Bayes Theorem and assumes that the influence of the features on the given classification is independent among other features, which differ from features that must fall into the specific category in other classifiers. Since this method has some positive characteristics, like relative high classification accuracy and high speed, it is relatively employed more on huge-capacity data processing and computation.

Artificial Neural Network (ANN or simply NN) (Multilayer Perceptron) [6, 12, 18, 38, 24] is inspired from human nervous system. Due to its nonlinear characteristic, it is proved that NN can normally render more accurate results compared with other single classifiers/ algorithms, especially when dealing with large database with high complexity or dimensionality.

Decision trees (DT) [55, 57, 56, 9, 11, 65], which has been widely used in many data mining applications as a supervised classifier. In machine learning, decision tree represents the mapping relationship between the object values and object properties. In current research, some researchers attempted to establish the rule-base classification systems through using keywords extracted from financial domain knowledge. In addition, when many decision

trees are combined together in order to improve the accuracy of classification, such method is known as the random forest (RF) classifier [43].

Majority Voting [43] refers to the use of multiple classifiers (which can be different types and setting), forming an ensemble to make classification prediction. The final output label is the one who got the majority vote in the ensemble of classifiers.

Waikato Environment for Knowledge Analysis (WEKA)

Waikato Environment for Knowledge Analysis (WEKA) [28] developed by the University of Waikato is used for most implementations related to machine learning in such experiment, including but not limited to the use of pattern recognition, features selections and evaluation. According to the initial announcement of its inventor group [28], The Waikato Environment for Knowledge Analysis (WEKA) came about through the perceived need for a unified workbench that would allow researchers to have easy access to the state-of-the-art techniques in machine learning.

Parameter Settings

The default parameter settings of the above five machine learning classifiers in WEKA are used in this experiment.

4.1.4 Other Experiment Design

Some details of experiment design are as follows:

First, multi-fold Cross-Validation has been applied for all experiments in this study in order to generate more reliable results avoiding bias as much as possible. On the basis of empirical results generated by [51], 10-fold cross-validation method is the relative optimal solution that keeps the balance between the calculation time and accuracy. Therefore, in this study, for all five machine learning algorithms, 10-fold Cross-Validation has been used.

Second, in order to overcome the intrinsic defects of each classifier for classification, the combination of all the above four classifiers in association with majority voting [62] scheme will be applied in this study. In comparison with results obtained by different classifier separately, we would like to improve the prediction accuracy and reduce the error variance in the meantime. Here, the majority voting scheme refers to the label that is given by majority classifiers can be output as the results.

Table 4.1: A summary of the 10-fold cross-validation result evaluated in Accuracy (%).

	Numerical Features Only	Numerical and Textual Features
Naive Bayes	89.4737	**90.3509**
Multilayer Perceptron (ANN)	88.5965	**92.1053**
Decision Tree (J48)	82.4561	**93.8596**
Random Forest	92.1053	**92.9825**
Majority Voting	89.4737	**94.7368**

Table 4.2: A summary of the 10-fold cross-validation result evaluated in F-measure.

	Numerical Features Only	Numerical and Textual Features
Naive Bayes	0.894	**0.902**
Multilayer Perceptron (ANN)	0.884	**0.921**
Decision Tree (J48)	0.820	**0.938**
Random Forest	0.920	**0.929**
Majority Voting	0.893	**0.947**

4.1.5 Performance Evaluation

Four commonly applied measurements in machine learning are utilized in this study to evaluate the performance of predictive models constructed by machine learning classifiers, as follows: Overall Accuracy, F-measure, ROC Area and Type I Error Rate [11, 58, 16, 68, 42, 34, 47, 41]. For the definition of these metrics, please refer to Appendix D for details.

4.2 Experiment Results and Analysis

Two datasets have been implemented in this research. For set A, only numerical features extracted from financial ratios have been used. For set B, the numerical features extracted from financial ratios and textual features generated through proposed concept-based textual feature extraction based on cTP2K have been combined together for this experiment. The detailed results about these two datasets in this experiment are displayed in Table 4.1, Table 4.2, Table 4.3 and Table 4.4.

Table 4.3: A summary of the 10-fold cross-validation result evaluated in ROC Area.

	Numerical Features Only	Numerical and Textual Features
Naive Bayes	0.986	**0.988**
Multilayer Perceptron (ANN)	0.960	**0.974**
Decision Tree (J48)	0.866	**0.949**
Random Forest	0.990	**0.991**
Majority Voting	0.985	**0.993**

Table 4.4: A summary of the 10-fold cross-validation result evaluated in Type I Error Rate

	Numerical Features Only	Numerical and Textual Features
Naive Bayes	0.053	**0.048**
Multilayer Perceptron (ANN)	0.057	**0.039**
Decision Tree (J48)	0.088	**0.031**
Random Forest	0.039	**0.035**
Majority Voting	0.053	**0.026**

4.2.1 Performance of Using Numerical Features Solely

Table 4.1 shows that the highest accuracy 92.10 was rendered through Random Forest if only Numerical Features applied. And the lowest accuracy is 82.46 generated by Decision Tree (J48). The accuracy through majority voting scheme reaches 89.50, and the average accuracy of above five algorithm is 88.42. The average Type I error is near 6%.

Compared with peers' achievements, these accuracy has no obvious competitiveness. For instance, as opposite to 82.46%, Olson et al. [51] achieved 91.4% accuracy through Decision Tree (J48) by using the WEKA similarly and selecting the similar number of features simultaneously. Meanwhile, Jardin [20] achieved 94.03% accuracy through neural networks (Multilayer Perceptron) with the similar number of features selection, in contrast to 88.60% in this experiment.

Nevertheless, it is notable that both of the above two literature applied bankruptcy companies and non-bankruptcy companies as binary classes samples in research. Correspondingly, the research samples in this experiment are three classes samples. Considering their actual financial status, we find that if only financial ratios professionally applied as the features, the discriminability of the same financial ratio in former bankruptcy and non-bankruptcy binary-level samples is obviously higher than the multi-level samples proposed in this research. Therefore, the decline of accuracy is reasonable in a similar experiment environment.

Suppose that a company has all its financial indicators in extreme terrible conditions, whether so-called binary prediction is still meaningful. In comparison with typical binary bankruptcy prediction model, this multi-level bankruptcy risk assessment system narrowed the scope between the upper sample class and lower sample class in association with inter-layer class, and generated the acceptable accuracy of assessment without extreme samples as before.

4.2.2 Performance of Using Combined Numerical Features and TP2K Textual Features

Table 4.1, Table 4.2, Table 4.3 and Table 4.4 also provide the summary of assessment performance results when we combine Numerical Features and cTP2K Textual Features. The highest accuracy obtained is 94.74% by Majority Voting in association with 0.026 type I error. At the same time, even the lowest overall accuracy rendered by Naive Bayes still exceeds 90%, which is 90.35%. The average accuracy of the performance exceeded 92.80% with 0.036 average type I error. From above tables, it is clear that the accuracy of Multilayer Perceptron, Decision Tree (J48) and majority voting have been greatly improved in comparison with Set A experiment. It is observed in Table 4.1, Table 4.2, Table 4.3 and Table 4.4 that Set B, combined with Textual Features extracted from TP2K, could prominently improve the classification accuracy when we compare the results with Set A. It has been demonstrated that the use of TP2K Textual Features is able to improve the accuracy of the proposed bankruptcy risk assessment system in similar experimental environments.

From the practical viewpoints, the followings are some underlying reasons to explain why it is able to enhance prediction results when cTP2K textual features from Corporate's Annual Reports were used.

First, the numbers of financial ratios could be manipulated under the accounting standards. In some cases, the companies close to bankruptcy tend to make their common financial ratios look good. Such acts will create another research challenge in financial fraud prediction [8, 52]. As a result, some common financial ratios are not discriminatbility as it was thought to be for identifying the upper sample class as well as the lower sample class related to bankruptcy assessment, like R1 the Current Ratio. However, when cTP2K textual features were extracted from the description of the previous performance of the company and the future expectations based on current performance, such textual information, as supplementary, could help to resolve such problems.

Second, numerical features transferred from financial ratios contain some intrinsic defects.

For instance, several financial ratios may depict the same characteristic of a company, e.g.in Fig.4.2, R1- Current Ratio and R2- Quick Ratio, and that is why R2 was abandoned in this experiment. Although, for machine learning, in general, providing more features may obtain relatively higher accuracy, unless inputted features are redundant. In the bankruptcy assessment problem ,this may also introduce biases. Thus, this is a concern and hence a balance is needed . Adding additional legally bound cTP2K textual features in the corporate annual reports can complement the utilization of numerical features and can also generate more useful assessment result.

Thirdly, based on the experimental observation in association with finance domain knowledge, there is a complementary relationship between the numerical features and cTP2K textual features in our experiment. For example, if an enterprise invests a large amount of fixed assets to enter a new industry, then some financial ratios, like R1/5/6 in Fig.4.2, could not be profitable and acceptable, and will indicate high-risk. Such cases always happen in practice in corporate performance assessment. However, in this situation, if the number of employee, which is T5 in Fig.4.3 used and extracted in this experiment, will increase. It can be considered that the financial conditions of the company may be better than it looks as it is unreasonable to hire more employees when it is under real financial distress.

In comparison with the existed research, we have found a model which analyzed the financial event with the combination of statistics-based textual features (high-frequency words) and numerical features (financial ratios) as reported in Cecchini et al. [8]. It renders 84% accurate prediction results over 156 sample data. In addition to this one, Shirata et al. [60] proposed a model to identify some specific key words to assist distinction of bankruptcy and non-bankruptcy of a company and achieved a prediction rate of 83% accuracy over 180 sample companies. Geng et al. [24] rendered 84% accurate prediction results among 107 sample data. In comparison, we found that the accuracy of our proposed bankruptcy risk assessment system cP2K has outperformed all the research results of our counterparts in our literature review.

In practice, the use of textual features is still confronting a universal problem, that is: how to quantify and interpret them with professional financial perspectives. Referring to existing literature, we still find this as an open question. Under such circumstances, in this experiment, there are three different valuations: numerical valuation, binary valuation and descriptive valuation to binarize the extracted textual features. From our experimental results between Set A and Set B, we observed that the binarized textual features maintain their discrimination in association with professional financial interpretation through this combined valuations mechanism. The additional three different valuation textual features also enhance the prediction. Before generally accepted criteria are proposed to evaluate textual features in finance, the combined valuations mechanism of cTP2K is an appropriate

substitute solution.

In brief, the results listed in Table 4.1, Table 4.2, Table 4.3 and Table 4.4 in association with above analysis can be considered as evidence to demonstrate the superiority of using cTP2K Textual Features.

4.2.3 Two Actual Samples Using Textual Features Extracted from cTP2K

First, there is a company RENTIAN TECH (Listed Code: 00885) which could serve as a sample of constituent stock companies under the Composite Mid-Small Cap Index (standards-risk) in our experiment. Due to its relatively negative financial ratios, it was easier to be classified as relatively high-risk company.

However, in its WTP Table, a low-frequency word acquisition (occurring only five times) was found and located in its directors report section of the 2015 annual report. The context containing such term was found through this key word. It was stated as follows: the Group made 5 additional acquisitions during the Year. This information can be considered as supplementary to show the positive future prospects of the company. At the same time, further investigation of its negative financial ratios could be conducted. Hence, in set B experiment, this sample company has been correctly classified when both numerical and textual features are used. If statistical frequency is applied to extract textual features, this key word will be ignored definitely. However, as presented in the walk-through of proposed cTP2K in Appendix A, it was found from the context and validated as a piece of crucial information, for predicting bankruptcy assessment.

This sample is a piece of evidence to show that cPT2K is able to overcome the low pattern frequency WTPs, a problem which is very common in current natural language processing based on statistics-based textual feature extraction. Based on the technology developed by Professor A.K.C. Wong [66], TP2K is built upon the intrinsic association of long digit sequences that reflects strong statistical association of multilevel relations within and between words to make up word, term and phrase (WTP) patterns. It is based on the statistical residuals (not ad hoc) and much more definitive and stable in identifying significant WTPs. Most important is TP2K is based on the notion that the default model of significant patterns is generated from random digit associations delimited by hash code. In another word, a WTP is statistical significant if the association of the digits in the sequence of C-code deviate from the default random digit model rather than random word model. Therefore, TP2K will not be affected by low frequency WTPs.

Second, before this experiment, Remuneration Policy is known as a subtitle in Management Discussion and Analysis section of corporate annul report. Through the cTP2K in association with accumulated Universal AVPs Table, several key attributes, such as Provident Fund Scheme, Medical Insurance, Retirement Benefit Scheme and Discretionary Bonuses were found with their value located within their embedding context confirming their relevancy in our experiment. It is clear that all these four attributes are commonly involved within Remuneration Policy based on 114 selected samples, and can be regarded as co-occurrence elements from a pattern discovery perspective. Without self-adaptive universal AVPs Table and local Domain Knowledge Base (DKB), it is really difficult to summarize and define such new knowledge in association with finance domain interpretation through actual samples.

Compared to the fixed predefined knowledge base applied in current research, the self-adaptive extendable Domain Knowledge Base (DKB) in cTP2K is self-extended and adaptive automatically when a new content is added or obtained with individual preferences or the expanding task environments and goals. As more and more data will be added to the system, cTP2K is supposed to become more and more intelligent as time goes by. This is the embodiment of advancement of proposed Concept-Based Extraction Textual Features based on TP2K (cTP2K) as widely applied statistics-based textual features are desired in bankruptcy assessment.

4.2.4 Some Practical Implications

Here, two significant points should be emphasized in association with the experiment results as listed in the results listed in Table 4.1 and Table 4.4.

First, in bankruptcy risk assessment research, Type I Error rate, which is also known as false positive rate, is the crucial measurement indicator but it is habitually misused. The underlying implication of its acceptance is that a relatively high-risk company is classified as an approximate risk-free company. Such an implication will definitely increase the possibility of enormous losses in actual investment. From the perspective of investment, safe investment is more important than missing an investment opportunity. However, the Type I Error rate is neglected by most of the current research. It is not usually released in the medium risk assessment or included in experimental results. Despite the lack of sufficient contradictory samples, in comparison with results generated from Set A and Set B through the majority voting schemes in our experiment, the average Type I Error rate can be reduced by more than 50% to 0.026 resulting in the improvement of our systems which surpass existing systems.

Second, in order to ensure that the assessment results can be interpreted professionally and practically, the domain knowledge must be introduced as the base to bridge the experimental result and the practical significance. In our experiment, machine learning algorithms and textual feature extraction are vital instruments to implement the proposed system according to the system design. Nevertheless, since the research of bankruptcy risk assessment is coming from the finance field, some financial or accounting knowledge should be incorporated and used in evaluating the numerical features as well as the proposed concept-based textual features. Incorporating expert judgment and final confirmation from the statistical and textual results is the essence of this bankruptcy risk assessment system.

Third, cTP2K is designed to be semi-automatic and interactive. It could bridge the gap between machine and human, and hence its extracted/mined textual features are more accurately and succinctly procured from corporate financial annual reports to allow domain expert interpretation and validation as well as professional practice with limited human participation. For professional practice, it is crucial to involve responsible personnel but with more succinct, comprehensive verifiable assistance from the machine intelligence to give greater confidence and credibility for the final decisionmaking. The new trend in machine intelligence is to bring machine closer to human but not to separate them as both parties will make each other much more intelligent and reliable as time goes by.

4.2.5 Summary

In summary, the result of this dissertation, data Set B, which combined the textual features extracted from cTP2K and numerical features derived from the annual reports together, could prominently improve the classification accuracy of bankruptcy assessment in comparison to the results obtained from data Set A, which utilize the numerical features exclusively. Therefore, the usefulness and practical significance of the complementary textual features extracted from proposed cTP2K have been demonstrated in real world scenarios.

Moreover, in comparison with the typical binary bankruptcy and non-bankruptcy samples, the scope between the upper class and lower class used in this experiment is relatively narrow. This resulted in the reduction of the statistical significance in numerical features obtained from each sample class. However, when combined with the cTP2K textual features such hurdles can be overcome. cTP2K was able to produce assessment results which are superior to others with a relatively narrow Type I Error rate. Therefore, the proposed multi-level bankruptcy risk assessment through cTP2K performed better and thus could replace the traditional binary prediction model.

Finally, the users need to oversee the whole procedure, from feature selection and valua-

tion to the machine learning phase. Although, some domain knowledge has been predefined in the existing domain knowledge base (like a pre-inputted attribute list) and used as assessment criteria, an interface with the user can help to add or change this list and criteria in compliance with his or her task objectives. In short, the concept-based textual feature extraction method cTP2K adopts a semi-supervised approach which ensures the interpretability and creditability of the textual features that are used and extracted. All discovered patterns and the contexts containing them can be located and displayed simultaneously in human-computer interaction platform of cTP2K, which is flexible, user-friendly and convenient for users to understand the underlying meaning of the discovered patterns and their usage. Hence, from the entire experiment procedure, there is an appropriate balance of efficiency, accuracy and user interpretation.

No.	Category	Financial Ratio Name
R1	Liquidity-Solvency	Current Ratio
R2	Liquidity-Solvency	Quick Ratio
R3	Liquidity-Solvency	Cash Flows from Operations to Current Liabilities Ratio
R4	Liquidity-Solvency	Cash Flows from Operations to Debt Ratio
R5	Liquidity-Solvency	Long-Term Debt to Equity Ratio
R6	Liquidity-Solvency	Debt to Equity Ratio
R7	Capital Structure	Interest Coverage Ratio
R8	Capital Structure	Debt to Assets Ratio
R9	Capital Structure	Equity Multiplier (EM)
R10	Capital Structure	Debt to Equity Ratio
R11	Capital Structure	Current Assets to Total Assets Ratio (CATA)
R12	Profitability-Growth	Gross Profit Margin
R13	Profitability-Growth	Net Profit Margin
R14	Profitability-Growth	Return on Equity (ROE)
R15	Profitability-Growth	Return on Invested Capital (ROIC)
R16	Profitability-Growth	Operating Profit Margin Ratio
R17	Profitability-Growth	Cash Flows from Operations to Net Income Ratio
R18	Profitability-Growth	Operation Income Growth Rate
R19	Operation Capacity	Inventory Turnover Rate
R20	Operation Capacity	Current Assets Turnover Rate
R21	Operation Capacity	Total Assets Turnover Rate
R22	Investment Value	Price-Earnings Ratio
R23	Investment Value	Price-Net Book Value Ratio
R24	Investment Value	Price-Sales Ratio
R25	Investment Value	Dividend Yield

Figure 4.2: The list of applied numerical features and corresponding financial ratio names

Numeber	Textual Feature Name	Representative Correspodning Attribute Names
T1	Risks / Uncentaties	e.g. contingent liabilities / discountinued operations
T2	Prospects / Outlooks	e.g. public float / acquisition / fund raising
T3	Remunaration / Emolument	e.g. discretionary bonuses / house allowance
T4	Exchange Expousre	e.g. currency risk / hedging tool
T5	Employee / Human Resource	employee
T6	Dividend	dividend
T7	Donation	donation
T8	Incentive Policy	e.g. share option scheme
T9	Pledge of Assets / Deposit	e.g. pledge of assets / pledge of bank deposit
T10	Major Customers / Suppliers	e.g. major customer / major supplier

Figure 4.3: The list of applied textual features and corresponding attribute names

Chapter 5

Conclusion

In this research, a method called concept-based textual feature extraction based on TP2K, cTP2K in abbreviation, is proposed. It has been successfully tested for semi-automatic extraction of AVPs (attribute names and their values) from text data of official corporate annual financial reports. A multi-level bankruptcy risk assessment system in association with its concept-based textual feature extraction (cTP2K) has also been proposed and demonstrated. This system can be subdivided into two parts: Feature Extraction Phase and Machine Learning Phase. By integrating the textual features (from extracted AVPs obtained from the text in corporate annual reports) with numerical features (financial ratios obtained from corporate annual reports) for predictive analysis, the experimental results demonstrated that prediction models constructed from both TP2K texture features and numeric features were superior to prediction models constructed only from numeric features. In our experiments under 10-fold cross-validation, the proposed system rendered more accurate classification results (almost reaching 94.74%) through combining numeric and cTP2K textural features together with 2.60% type-1 error rate through majority voting scheme. Based on these results, we list our major contributions as follows:

First, to the best of my knowledge, multi-level bankruptcy risk assessment (risk-free, standard and high-risk) is the first time being proposed in the field of finance, comparing to the traditional binary bankruptcy risk assessment (risk-free and high-risk). It is a more realistic setting and we believe this would move the bankruptcy risk assessment research closer to actual practice.

Second, the good classification results above have demonstrated that useful and appropriate textual features can be extracted through TP2K particularly in parts of the annual report in a semi-automatic manner for machine learning modeling. The use of concept-

based textual feature extraction based on TP2K can render a fast, more robust and less biased feature search and compilation process with statistical support for experts to acquire and organize their rules/knowledge in a comprehensive/ and analytical manner.

Third, in association with actual samples in experiments, the preliminary design of the self-adaptive Domain Knowledge Base (DKB), a component module in concept-based textual feature extraction based on TP2K, has also demonstrated its practicability to support user or experts in practice, and can be regarded as a future extension of bankruptcy risk assessment research. It is anticipated that the effectiveness of this automated machine intelligence process and domain knowledge built-up will increase as more knowledge is acquired and organized so as to take over more of the human effort in the long run in the future.

There are also suggestions about the future work related to this research:

First, the research about bankruptcy risk assessment is derived from real-world professional practice. The proposed multi-level bankruptcy risk assessment in this research suggests that the system should have the following three characteristics: transparency, traceability, and extensibility. Therefore, multi-level bankruptcy risk assessment system in association with concept-based textual feature extraction based on TP2K could bridge the gap between machine and human intelligence, and the output should allow industry experts to interpret and analyze. Future systems should have these characteristics.

Second, in this study, cTP2K is the primary tool to discover and locate concept-based textual features from the unstructured context, and disclose them in a succinct and comprehensive way for in-depth classification and assessment by machine and users. Its application can further be expanded in the future. For example, the attribute names at present are pre-inputted from the section titles of annual reports. In the future, their more generic patterns could be extracted from cTP2K and applied to enrich the domain knowledge base (DKB) so as to provide more comprehensive attributes in financial practice.

Third, to investigate the scalability before real-world practice, a larger scale of experiments shall be designed to study if the prediction performance could be further improved through the proposed multi-level bankruptcy risk assessment system.

In short, the proposed multi-level bankruptcy risk assessment system can potentially be used to solve real-world bankruptcy risk assessment problems with professional financial interpretation. It should also be noted that although this study focuses on problems in finance, there is no presumption that would hinder its effective application in other fields. From another perspective, the concept-based textual feature extraction that is based on TP2K could be leveraged with the accumulated practical experience in any other disciplinary.

References

[1] Vineet Agarwal and Richard Taffler. Comparing the performance of market-based and accounting-based bankruptcy prediction models. *Journal of Banking & Finance*, 32(8):1541–1551, 2008.

[2] Edward I Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4):589–609, 1968.

[3] Martin Aruldoss, Miranda Lakshmi Travis, and V Prasanna Venkatesan. A reference model for business intelligence to predict bankruptcy. *Journal of Enterprise Information Management*, 28(2):186–217, 2015.

[4] Mateusz Baryła, Barbara Pawełek, and Józef Pociecha. Selection of balanced structure samples in corporate bankruptcy prediction. In *Analysis of Large and Complex Data*, pages 345–355. Springer, 2016.

[5] William H Beaver. Financial ratios as predictors of failure. *Journal of accounting research*, pages 71–111, 1966.

[6] Indranil Bose and Raktim Pal. Predicting the survival or failure of click-and-mortar corporations: A knowledge discovery approach. *European Journal of Operational Research*, 174(2):959–982, 2006.

[7] Carlo Caserio, Delio Panaro, and Sara Trucco. Management discussion and analysis in the us financial companies: A data mining analysis. In *Strengthening Information and Control Systems*, pages 43–57. Springer, 2016.

[8] Mark Cecchini, Haldun Aytug, Gary J Koehler, and Praveen Pathak. Making words work: Using financial text as a predictor of financial events. *Decision Support Systems*, 50(1):164–175, 2010.

[9] Samuel WK Chan and James Franklin. A text-based decision support system for financial sequence prediction. *Decision Support Systems*, 52(1):189–198, 2011.

[10] Te-Min Chang, Ming-Fu Hsu, Guo-Hsin Hu, and Keng-Pei Lin. Salient corporate performance forecasting based on financial and textual information. In *Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on*, pages 000959–000964. IEEE, 2016.

[11] Mu-Yen Chen. Predicting corporate financial distress based on integration of decision tree classification and logistic regression. *Expert Systems with Applications*, 38(9):11261–11272, 2011.

[12] Wei-Sen Chen and Yin-Kuan Du. Using neural networks and data mining techniques for the financial distress prediction model. *Expert Systems with Applications*, 36(2):4075–4086, 2009.

[13] Ching-Hsue Cheng and Chia-Pang Chan. An attribute selection based classifier to predict financial distress. In *Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 2016 12th International Conference on*, pages 1119–1124. IEEE, 2016.

[14] Anja Cielen, Ludo Peeters, and Koen Vanhoof. Bankruptcy prediction using a data envelopment analysis. *European Journal of Operational Research*, 154(2):526–532, 2004.

[15] Loredana Cultrera and Bauweraerts Jonathan. Exploring corporate bankruptcy in belgian private firms. *International Journal of Economics and Finance*, 9(3):108, 2017.

[16] Dursun Delen, Cemil Kuzey, and Ali Uyar. Measuring firm performance using financial ratios: A decision tree approach. *Expert Systems with Applications*, 40(10):3970–3983, 2013.

[17] Umberto Dellepiane, Michele Di Marcantonio, Enrico Laghi, and Stefania Renzi. Bankruptcy prediction using support vector machines and feature selection during the recent financial crisis. *International Journal of Economics and Finance*, 7(8):182, 2015.

[18] Satyajit Dhar, Tuhin Mukherjee, and Arnab Kumar Ghoshal. Performance evaluation of neural network approach in financial prediction: Evidence from indian market.

In *Communication and Computational Intelligence (INCOCCI), 2010 International Conference on*, pages 597–602. IEEE, 2010.

[19] Augustinos I Dimitras, Stelios H Zanakis, and Constantin Zopounidis. A survey of business failures with an emphasis on prediction methods and industrial applications. *European Journal of Operational Research*, 90(3):487–513, 1996.

[20] Philippe Du Jardin. Predicting bankruptcy using neural networks and other classification methods: The influence of variable selection techniques on model accuracy. *Neurocomputing*, 73(10):2047–2060, 2010.

[21] Kenneth Fisher. *Super stocks*. McGraw Hill Professional, 2007.

[22] George Foster. *Financial Statement Analysis, 2/e*. Pearson Education India, 1986.

[23] Christopher Fox. A stop list for general text. In *Acm sigir forum*, volume 24, pages 19–21. ACM, 1989.

[24] Ruibin Geng, Indranil Bose, and Xi Chen. Prediction of financial distress: An empirical study of listed chinese companies using data mining. *European Journal of Operational Research*, 241(1):236–247, 2015.

[25] Glen L Gray and Roger S Debreceny. A taxonomy to guide research on the application of data mining to fraud detection in financial statement audits. *International Journal of Accounting Information Systems*, 15(4):357–380, 2014.

[26] Sven S Groth and Jan Muntermann. An intraday market risk management approach based on textual analysis. *Decision Support Systems*, 50(4):680–691, 2011.

[27] Petr Hájek and Vladimír Olej. Word categorization of corporate annual reports for bankruptcy prediction by machine learning methods. In *International Conference on Text, Speech, and Dialogue*, pages 122–130. Springer, 2015.

[28] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[29] Sachio Hirokawa, Takahiro Baba, and Tetsuya Nakatoh. Text mining of bankruptcy information using formal concept analysis. In *Awareness Science and Technology (iCAST), 2011 3rd International Conference on*, pages 527–532. IEEE, 2011.

[30] Hongkyu Jo, Ingoo Han, and Hoonyoung Lee. Bankruptcy prediction using case-based reasoning, neural networks, and discriminant analysis. *Expert Systems with Applications*, 13(2):97–108, 1997.

[31] Nam-ok Jo and Kyung-shik Shin. Bankruptcy prediction modeling using qualitative information based on big data analytics. *??????*, 22(2):33–56, 2016.

[32] Hyun-Jung Kim, Nam-Ok Jo, and Kyung-Shik Shin. Optimization of cluster-based evolutionary undersampling for the artificial neural networks in corporate bankruptcy prediction. *Expert Systems with Applications*, 59:226–234, 2016.

[33] Myoung-Jong Kim and Ingoo Han. The discovery of experts' decision rules from qualitative bankruptcy data using genetic algorithms. *Expert Systems with Applications*, 25(4):637–646, 2003.

[34] Efstathios Kirkos. Assessing methodologies for intelligent bankruptcy prediction. *Artificial Intelligence Review*, pages 1–41, 2015.

[35] Antonina Kloptchenko, Tomas Eklund, Jonas Karlsson, Barbro Back, Hannu Vanharanta, and Ari Visa. Combining data and text mining techniques for analysing financial reports. *Intelligent systems in accounting, finance and management*, 12(1):29–41, 2004.

[36] B Shravan Kumar and Vadlamani Ravi. A survey of the applications of text mining in financial domain. *Knowledge-Based Systems*, 114:128–147, 2016.

[37] Salim Lahmiri. Features selection, data mining and finacial risk classification: a comparative study. *Intelligent Systems in Accounting, Finance and Management*, 23(4):265–275, 2016.

[38] Sangjae Lee and Wu Sung Choi. A multi-industry bankruptcy prediction model using back-propagation neural network and multivariate discriminant analysis. *Expert Systems with Applications*, 40(8):2941–2946, 2013.

[39] Tsun-Siou Lee and Yin-Hua Yeh. Corporate governance and financial distress: Evidence from taiwan. *Corporate governance: An international review*, 12(3):378–388, 2004.

[40] Feng Li. The information content of forward-looking statements in corporate filingsa naïve bayesian machine learning approach. *Journal of Accounting Research*, 48(5):1049–1102, 2010.

[41] Deron Liang, Chia-Chi Lu, Chih-Fong Tsai, and Guan-An Shih. Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European Journal of Operational Research*, 252(2):561–572, 2016.

[42] Deron Liang, Chih-Fong Tsai, and Hsin-Ting Wu. The effect of feature selection on financial distress prediction. *Knowledge-Based Systems*, 73:289–297, 2015.

[43] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[44] Feng-Yi Lin, Deron Liang, and Wing-Sang Chu. The role of non-financial features related to corporate governance in business crisis prediction. *Journal of Marine Science and Technology*, 18(4):504–513, 2010.

[45] Fengyi Lin, Deron Liang, Ching-Chiang Yeh, and Jui-Chieh Huang. Novel feature selection methods to financial distress prediction. *Expert Systems with Applications*, 41(5):2472–2483, 2014.

[46] Thomas E Mckee. Developing a bankruptcy prediction model via rough sets theory. *International Journal of Intelligent Systems in Accounting, Finance & Management*, 9(3):159–173, 2000.

[47] Kalyan Nagaraj and Amulyashree Sridhar. A predictive system for detection of bankruptcy using machine learning techniques. *arXiv preprint arXiv:1502.03601*, 2015.

[48] Arman Khadjeh Nassirtoussi, Saeed Aghabozorgi, Teh Ying Wah, and David Chek Ling Ngo. Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41(16):7653–7670, 2014.

[49] Timothy Nugent and Jochen L Leidner. Risk mining: company-risk identification from unstructured sources. In *Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on*, pages 1308–1311. IEEE, 2016.

[50] James A Ohlson. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of accounting research*, pages 109–131, 1980.

[51] David L Olson, Dursun Delen, and Yanyan Meng. Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems*, 52(2):464–473, 2012.

[52] G Ozdagoglu, A Ozdagoglu, Y Gumus, and G Kurt Gumus. The application of data mining techniques in manipulated financial statement classification: The case of turkey. *Journal of AI and Data Mining*, 5(1):67–77, 2017.

[53] Parag C Pendharkar. A threshold-varying artificial neural network approach for classification and its application to bankruptcy prediction problem. *Computers & Operations Research*, 32(10):2561–2582, 2005.

[54] Ivica Pervan and Tamara Kuvek. The relative importance of financial ratios and nonfinancial variables in predicting of insolvency. *Croatian Operational research review*, 4(1):187–197, 2013.

[55] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[56] Gil Rachlin, Mark Last, Dima Alberg, and Abraham Kandel. Admiral: A data mining based financial trading system. In *Computational Intelligence and Data Mining, 2007. CIDM 2007. IEEE Symposium on*, pages 720–725. IEEE, 2007.

[57] Manuel Filipe Santos, Paulo Cortez, José Pereira, and Helder Quintela. Corporate bankruptcy prediction using data mining techniques. *WIT transactions on information and communication technologies*, 37, 2006.

[58] Fu Shuen Shie, Mu-Yen Chen, and Yi-Shiuan Liu. Prediction of corporate financial distress: an application of the america banking industry. *Neural Computing and Applications*, 21(7):1687–1696, 2012.

[59] Kyung-Shik Shin and Yong-Joo Lee. A genetic algorithm application in bankruptcy prediction modeling. *Expert Systems with Applications*, 23(3):321–328, 2002.

[60] Cindy Yoshiko Shirata, Hironori Takeuchi, Shiho Ogino, and Hideo Watanabe. Extracting key phrases as predictors of corporate bankruptcy: Empirical analysis of annual reports by text mining. *Journal of Emerging Technologies in Accounting*, 8(1):31–44, 2011.

[61] Jie Sun and Hui Li. Data mining method for listed companies financial distress prediction. *Knowledge-Based Systems*, 21(1):1–5, 2008.

[62] Jie Sun and Hui Li. Listed companies financial distress prediction based on weighted majority voting combination of multiple classifiers. *Expert Systems with Applications*, 35(3):818–827, 2008.

[63] Jie Sun and Hui Li. Financial distress early warning based on group decision making. *Computers & Operations Research*, 36(3):885–906, 2009.

[64] Jie Sun, Hui Li, Qing-Hua Huang, and Kai-Yu He. Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. *Knowledge-Based Systems*, 57:41–56, 2014.

[65] Tien-Thanh Vu, Shu Chang, Quang Thuy Ha, and Nigel Collier. An experiment in integrating sentiment features for tech stock prediction in twitter. *Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data*, 2012.

[66] Andrew KC Wong, Dennis Zhuang, Gary CL Li, and En-Shiun Annie Lee. Discovery of delta closed patterns and noninduced patterns from sequences. *IEEE Transactions on Knowledge and Data Engineering*, 24(8):1408–1421, 2012.

[67] Zhi Xiao, Xianglei Yang, Ying Pang, and Xin Dang. The prediction for listed companies financial distress by using multiple prediction methods with rough set and dempster–shafer evidence theory. *Knowledge-Based Systems*, 26:196–206, 2012.

[68] Ligang Zhou. Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods. *Knowledge-Based Systems*, 41:16–25, 2013.

[69] Ligang Zhou, Kin Keung Lai, and Jerome Yen. Empirical models based on features ranking techniques for corporate financial distress prediction. *Computers & Mathematics with Applications*, 64(8):2484–2496, 2012.

[70] Ligang Zhou, Dong Lu, and Hamido Fujita. The performance of corporate financial distress prediction models with features selection guided by domain knowledge and data mining approaches. *Knowledge-Based Systems*, 85:52–61, 2015.

APPENDICES

Appendix A

A walk-through of a demonstration example of Applying to c-TP2K in Bankruptcy Risk Assessment

The sample text data used in this walk-through comes from the Director' s Report and Management Discussion and Analysis section of 2015 Annual Report of RENTIAN TECH (listed code: 00885). The procedure of this walk-through is described in steps as listed in Fig. A.1. The following screenshots show part of the data/ database used in this experiment.

Inputted Original Text

The Inputted Original Text, in Fig. A.2, is extracted from corresponding corporate annual report directly; therefore, its text content is complicated and chaotic, and further organization is needed here.

Step (1) Transcribe the Original Text into ω-Text

The Inputted Original Text will be arranged such that all sentences are represented line by line through cTP2K automatically. We refer such text file as ω-Text, as shown in Fig. A.3.

Number of Step	System Module	Function Description
1	Data Pre-Processing System	Transfer Original Text into ω-Text
2		Transfer ω-Text into W-Text
3		Transfer W-Text into C-Text
4	Patterns Discovery System	From C-Text to C-Pattern
5		From C-Pattern to WTP Table (Word-Tern-Phrase)
6	Pattern Embedded Context Finding & Locating System	Find & Locate Context through WTP Table
7		Find & Locate Context through Domain Knowledge
8	Attribute Name and Value Pairs (AVPs) Extraction System	Validate Attribute Name
9		Validate Attribute Value
10		Attribute Name and Value Pairs (AVPs) Table
11	Self-Extended Domain Knowledge Base (DKB) System	Construct Local Domain Knowledge Base (DKB)

Figure A.1: The Procedure of Walk-through

Step (2) Transcribe ω-Text into W-Text

The cTP2K will automatically remove numerals and common words in accordance with a pre-defined set (a list of common words [23], like articles, prepositions, numerals etc., which do not contribute to make up the patterns relevant to what we intend to discover). Such a list is flexible and dynamically adaptable. We refer such step as: From ω-Text to W-Text, as shown in Fig. A.4.

It is notable here that empty sentences in the W-Text are removed. In order to keep the same sentence number (assigned line by line as an address) of specific sentence in ω-Text and W-Text, period will still be placed there to mark the empty lines and keep their positions.

Here, for example, the sentence in the first line of text file will be assigned s_1 as an address automatically.

Step (3) Translate the W-Text into C-Text

All uncommon word in W-Text will be converted into unique 6-digit hash code to form a sequence made up of 6-digit codes as word units with a spacing between. We refer to this text as C-Text, as shown in Fig. A.5.

Step (4) From C-Text to C-Pattern

C-Text is inputted as a text file to generate C-Patterns by clicking the functional key pattern discovery in the graphical user interface(GUI) of Pattern Discovery System developed by Dr. Wong and his team [66], as shown in Fig. A.6. Then the functional key save file can be used to output C-Patterns in csv format. The parameter setting of this system can be default or determined individually by user in the enclosed red box in Fig. A.6.

It is notable that there are two parameters here, which are minimum occurrence and delta. A pattern is extracted from the input text if its number of occurrences is larger than the minimum occurrence. This set of patterns are then compressed by an algorithm proposed by [66]. By setting delta to 1, all information contained by the patterns, i.e. 100%, would be retained, i.e. lossless compression. According to [66], by varying delta from 1 to 0, a certain compression rate can be achieved. For example, if delta is set to 0.8, about 80% of the information would be retained [66]. The default setting of minimum occurrence and delta, according to [66], are 5 and 0.8, respectively. Users are allowed to vary these parameter setting depending on their domain applications.

Fig. A.7 is a part of the generated C-Patterns made up of hash codes of the 2015 Annual Report of RENTIAN TECH (listed code: 00885) in csv format.

Step (5) From C-Patterns to WTP Table

All digit units with length less than 6-digit of the code-based patterns, will be abandoned since they are not parts of word patterns. After removal, all the retaining digit units of code-based patterns containing a sequence of Hash Code will be converted back to words. Then, a table of significant C-patterns discovered will be converted and organized into a word-based patterns table that makes up of significant Word, Term and Phrase patterns. We refer the table as a WTP Table, as shown in Fig. A.8.

Step (6) Find & Locate the Context containing any WTP in the WTP Table

The Step (6) can be divided into three parts: Part A, Part B and Part C.

For Part A, as shown in the red box in the Fig. A.9, there are five search conditions to examine and study the WTPs and their relevancy and usefulness from the WTP Table. They are: their number of occurrences, support, statistical significance, length and number of co-occurrences. In the GUI Fig. A.9, the top ten of most significant WTPs could be revealed and used as the search condition by clicking functional key search. Then the GUI will direct the user to Part B.

For Part B, as shown in the red box in the Fig. A.10, there is a list of patterns in the WTP table with top ten statistical significance WTP extracted for further review by the user. Here, we see that a significant pattern has relevancy to bankruptcy risk assessment, namely acquisition equity interests, with blue shade, has been found. Certainly, it could happen that the found patterns may have no relevancy to bankruptcy risk. If that is the case, a new search condition is needed. The user will move onto Part C.

For Part C, as shown in the red box in the Fig. A.11, after selecting pattern acquisition equity interests, its relevant information is presented and followed in the top right corner blocked in a red box contacting the WTP followed by 5, 5, 603604066.83, 19, which represent the number of occurrences, support, statistical significance, pattern length respectively [66]; and their contexts containing it as shown in the bottom window with selected significant pattern highlighted in yellow shade.

From the above Fig. A.11, it is clear that the extracted contexts from cTP2K is unabridged and comprehensible.

Step (7) Find & Locate Context through Domain Knowledge of Users

Here, as shown in Fig. A.12, the user types in any word(s) within the search window. Any word, e.g. profit before taxation, financial cost would be found if existing in the text and the contexts embracing them would be displayed in the bottom windows with the typed in WTP colored in yellow and gray shade respectively. These are the co-occurring WTPs within a sentence context. The user could understand how such domain knowledge is related to bankruptcy assessment, and then identify the attribute name (a relevant WTP) and its attribute value from the context.

Step (8) Validate Attribute Name

First, as shown in the red box in the Fig. A.13, in the GUI, the user can click functional key reload attribute to upload the predefined Local Attribute Names List. Then, the functional

key search attribute should be clicked in order to extract the attribute name automatically in each sentence.

Here, in the blue boxes in the Fig. A.13, it is being noted that the number in the beginning of each extracted attribute is corresponding with the number in the beginning of sentences, which makes users more easily to find the sentence where the attribute name was extracted from.

In order to facilitate the user to select the most concerned attribute name/s, users can also input any attribute name they prefer, as shown in Fig. A.14. cTP2K will check new inputted attribute name/s automatically, and see if it is already in the current local attribute names list. If not, it will be added automatically. This self-adaptive local attribute names list is also a part of the proposed Domain Knowledge Base (DKB) which can be expanded in accordance with the user's own preferences.

Step (9) Validate Attribute Values

Once, the attribute name/s and its corresponding contexts are found and located, the user can then review the context and determine the most appropriate value. Here, the attribute values are flexible to emphasize the individual research by the user based on his/her financial knowledge.

In order to avoid possible redundancy or repetition, as long as there is a complete pair of attribute name and value, the user can click the functional key output attribute to add it or them in the corresponding AVPs tables, as shown in Fig. A.15.

Step (10) Attribute Name and Value Pairs (AVPs) Table

The Fig. A.16 represents the tabulated Attribute Name and Value Pairs (AVPs) to form an AVPs Table for RENTIAN TECH (00885). From this table, the financial performance of this enterprise can be evaluated in a comprehensive manner. The textual features applied in this experiment are also selected from this table, e.g. employee, dividend and remuneration.

Here, this AVPs Table is self-adaptive. This means that all the AVPs extracted from this text file of RENTIAN TECH can be added into this table at any time.

Step (11) Construction of Local Domain Knowledge Base (DKB)

Fig. A.17 represents a self-extended AVPs Database. All the acquired AVPs tables will be added into this AVPs database as a part of the Domain Knowledge Base (DKB) in support of predictive analysis of corporate bankruptcy risk, e.g. from which one could find the most frequent attribute names. The AVPs in the red box in Fig. A.17 is part of AVPs of RENTIAN TECH (00885).

Furthermore, all inputted/predefined self-extended dictionaries/lists utilized in TP2K are self-extended, e.g. local attribute names list, are also utilized in the concept-based textual feature extraction based on TP2K (cTP2K) system. All these sets, lists and dictionaries can also be considered as a part of the extending Domain Knowledge Base (DKB).

PROSPECT

For Rentian Technology, 2015 marked a year of breakthrough and transformation into an integrated internet-of-things ("IoT") solution provider. The Group swiftly took ground in information flows, logistics, capital flows and other IoT businesses through expanding into data collection, virtual end-devices (endpoints), data analysis and transmission, cloud services and other fields. It is now equipped with all the tools that are fundamental to the provision of one-stop IoT solutions. Synergy from the acquisition of 5 upstream and downstream IoT companies (namely Shenzhen CNEOP Technology Company Limited* ("CNEOP"), Guangzhou Wealth-Depot Logistics Technology Company Limited* ("Wealth-Depot"), Shenzhen Hexicom Technologies Company Limited* ("Hexicom"), Fujian Start Computer Equipment Company Limited* ("FSCE") and an associated company namely Beijing Oriental Legend Maker Technology Limited* ("OLM")) is expected to enhance the Group's overall profitability by reducing procurement costs, speeding up supplies, improving quality and broadening clientele. Furthermore, there were also certain breakthroughs in overseas businesses in 2015.

Rentian Technology believes that the IoT market will continue to experience exponential growth in the coming years. As such, the Group will carry on marketing, product research and development as well as merger and acquisition in the IoT industry. According to an industry forecast conducted by International Data Corporation ("IDC") in February 2016, the global IoT market will expand from US$591.7 billion in 2014 to US$1,300.0 billion in 2019, representing an estimated compound annual growth rate (CAGR) of 17%. The number of installed endpoints is also estimated to increase from US$9.7 billion in 2014 to US$25.6 billion in 2019 and reach US$30.0 billion by 2020.

Figure A.2: Inputted Original Text

PROSPECT For Rentian Technology, 2015 marked a year of breakthrough and transformation into an ir
The Group swiftly took ground in information flows, logistics, capital flows and other IoT busine
It is now equipped with all the tools that are fundamental to the provision of one-stop IoT solut
Synergy from the acquisition of 5 upstream and downstream IoT companies (namely Shenzhen CNEOP Te
Furthermore, there were also certain breakthroughs in overseas businesses in 2015.
Rentian Technology believes that the IoT market will continue to experience exponential growth ir
As such, the Group will carry on marketing, product research and development as well as merger ar
According to an industry forecast conducted by International Data Corporation ("IDC") in Februa
The number of installed endpoints is also estimated to increase from US$9.7 billion in 2014 to US
Rentian Technology is also optimistic about the Chinese IoT market.
IDC anticipated that there will be 5.4 billion IoT endpoints in China by 2020, which means that 1
Meanwhile, the State Council of China issued a document entitled "Made in China 2025" in May 20
To achieve these goals, it will need to boost its Industry 4.0 or smart manufacturing capability,
Rentian Technology believes that the Group's current direction is in line with the long-term dev
Looking forward to 2016, Rentian Technology will invest in and develop the following aspects: 1.

Figure A.3: ω-Text

PROSPECT Rentian Technology marked year breakthrough transformation integrated int
Group swiftly took ground information flows logistics capital flows IoT businesses
equipped tools fundamental provision stop IoT solutions.
Synergy acquisition upstream downstream IoT companies Shenzhen CNEOP Technology Co
Furthermore certain breakthroughs overseas businesses.
Rentian Technology believes IoT market continue experience exponential growth comi
Group carry marketing product research development merger acquisition IoT industry
According industry forecast conducted International Data Corporation IDC February
number installed endpoints estimated increase billion billion reach billion.
Rentian Technology optimistic Chinese IoT market.
IDC anticipated billion IoT endpoints China means endpoints world installed China.
State Council China issued document entitled China forward Steps guiding principle
achieve goals need boost Industry smart manufacturing capability IoT industrial in
Rentian Technology believes Group current direction line long term development str
Looking forward Rentian Technology invest develop following aspects.

Figure A.4: W-Text

877580 810973 744044 761228 204893 916806 471337 986003 248833 364773 204462 991833 738705
629247 915616 965975 306695 582060 427045 814707 569064 427045 204462 157582 885608 476010
168991 845371 671557 550581 940994 204462 746938
659655 194733 840411 984322 204462 201819 574837 406203 744044 997309 964954 406203 931744
311353 242402 221101 752472 157582
810973 744044 398207 204462 761244 178119 783594 997381 308807 848417 651798
629247 931515 124422 641039 910875 466587 851482 194733 204462 694718
993858 694718 492955 708963 801390 476010 456040 204104 331714 146915 204462 761244 800090
903209 694330 215070 860380 359682 883745 883745 486675 883745
810973 744044 839213 243565 204462 761244
204104 700378 883745 204462 215070 231255 369422 215070 918802 694330 231255
957585 227649 231255 107403 905403 144117 231255 734597 961319 843033 862876 746542 231255
718065 526144 777302 422211 694718 749001 565406 424248 204462 111737 248833 218958 714154
810973 744044 398207 629247 153241 378431 721844 727612 956460 466587 910643 243565 200273
936195 734597 810973 744044 964245 904061 591537 751515

Figure A.5: C-Text

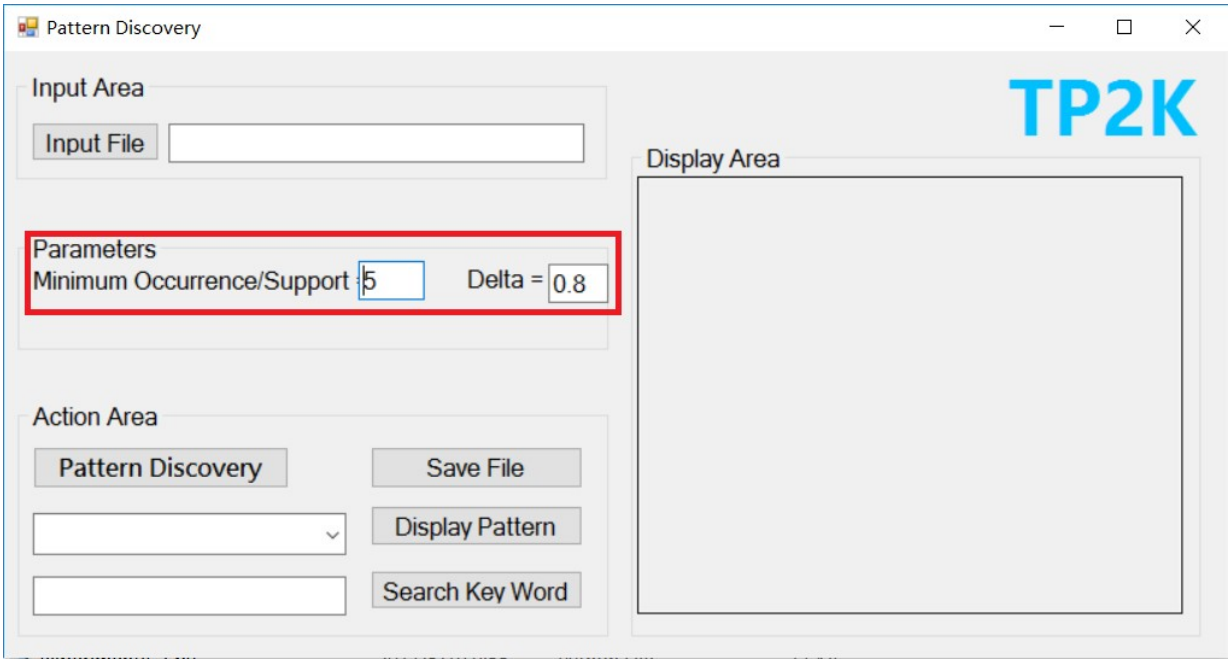


Figure A.6: Pattern Discovery System

Flexible Gap Pattern	# of occurrences	support	Sig	length	Positions
213762 787378 707963 773833 142436	7	7	2.77413E+15	33	s106: [28 60]*s111: [6
787378 707963 773833 142436	9	9	1.1998E+12	26	s106: [35 60]*s111: [7
194733 792293 859273	5	5	603604066.8	19	s36: [35 53]*s37: [21
707963 773833 142436	12	12	461437095.5	19	s94: [28 46]*s106: [42
204893 451962 389093	10	8	253825201.2	19	s32: [105 123]*s49: [2
204893 451962 389093	9	7	199887346	21	s42: [0 20]*s45: [0 20]
185153 224236 224236	5	3	177888439.6	18	s34: [63 80]*s34: [119
600031 830325 379749	5	4	68525912.61	21	s88: [21 41]*s133: [63
185153 224236	24	11	716840.37	13	s32: [49 61]*s32: [84
441805 172180	10	7	246006.25	13	s3: [91 103]*s3: [133
810973 744044	12	12	135220.93	14	s0: [7 20]*s5: [0 13]*s
451962 389093	14	11	130704.41	12	s32: [112 123]*s42: [7
608448 697821	6	5	97832.13	13	s51: [28 40]*s78: [70
554500 660936	6	4	76589.19	13	s47: [112 124]*s47: [1
608448 697821	9	7	75461.97	12	s51: [28 39]*s78: [70
997309 353227	6	6	65817.76	13	s94: [49 61]*s103: [49
600031 830325	8	7	55127.94	14	s34: [147 160]*s88: [2
306331 792995	5	5	51636.23	13	s38: [14 26]*s39: [49
132264 981275	5	5	23723.35	12	s67: [21 32]*s95: [91
554500 233802	5	4	20408.51	14	s32: [70 83]*s34: [0 1

Figure A.7: C-Pattern

Flexible Gap Pattern	# of Occurrences	Support	Sig	Length	Positions
set note consolidated financial statements	7	7	2774132633716920.33		s106: [28 60]*s111: [63
note consolidated financial statements	9	9	1199804381288.41	26	s106: [35 60]*s111: [70
acquisition equity interests	5	5	603604066.83	19	s36: [35 53]*s37: [21 3
consolidated financial statements	12	12	461437095.51	19	s94: [28 46]*s106: [42
year ended december	10	8	253825201.22	19	s32: [105 123]*s49: [21
year ended december	9	7	199887345.96	21	s42: [0 20]*s45: [0 20]
approximately million million	5	3	177888439.56	18	s34: [63 80]*s34: [119
share option scheme	5	4	68525912.61	21	s88: [21 41]*s133: [63
approximately million	24	11	716840.37	13	s32: [49 61]*s32: [84 9
wealth depot	10	7	246006.25	13	s3: [91 103]*s3: [133 1
rentian technology	12	12	135220.93	14	s0: [7 20]*s5: [0 13]*s
ended december	14	11	130704.41	12	s32: [112 123]*s42: [7
hong kong	6	5	97832.13	13	s51: [28 40]*s78: [70 8
profit guarantee	6	4	76589.19	13	s47: [112 124]*s47: [14
hong kong	9	7	75461.97	12	s51: [28 39]*s78: [70 8
company subsidiaries	6	6	65817.76	13	s94: [49 61]*s103: [49
share option	8	7	55127.94	14	s34: [147 160]*s88: [21
principally engaged	5	5	51636.23	13	s38: [14 26]*s39: [49 6
general meeting	5	5	23723.35	12	s67: [21 32]*s95: [91 1
profit taxation	5	4	20408.51	14	s32: [70 83]*s34: [0 13

Figure A.8: WTP Table

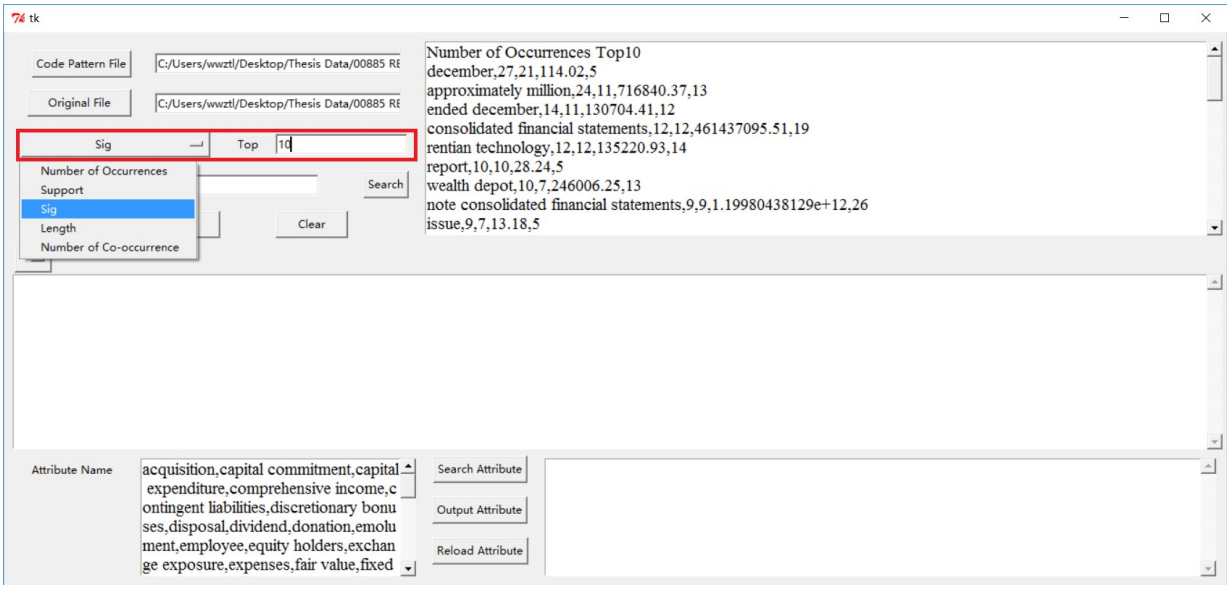


Figure A.9: Part A of Step (6)

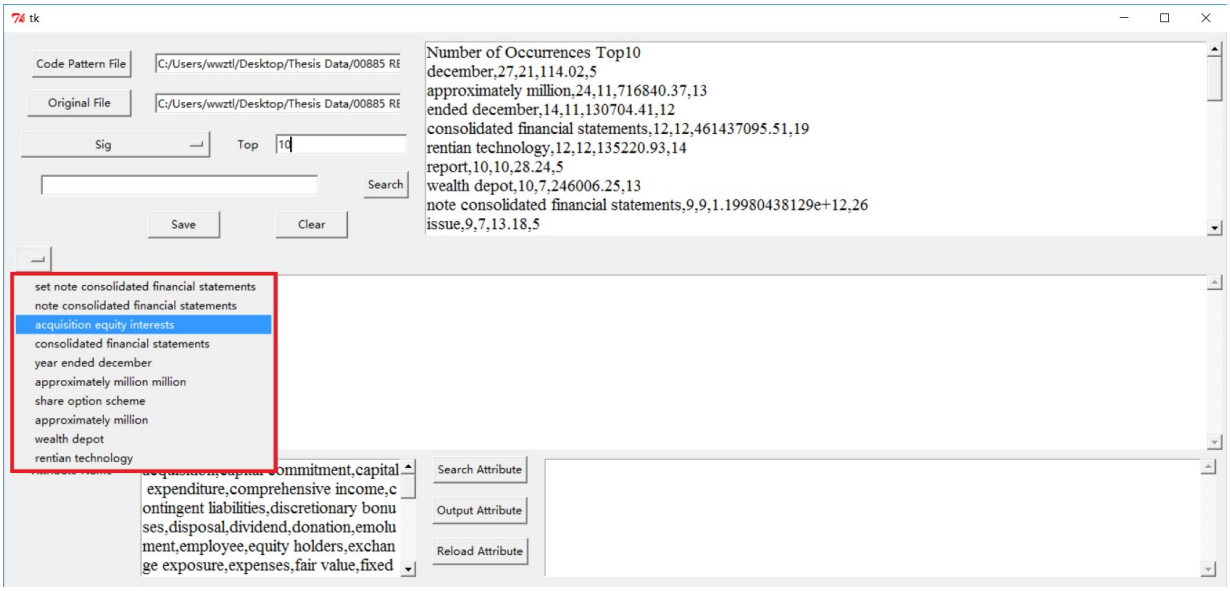


Figure A.10: Part B of Step (6)

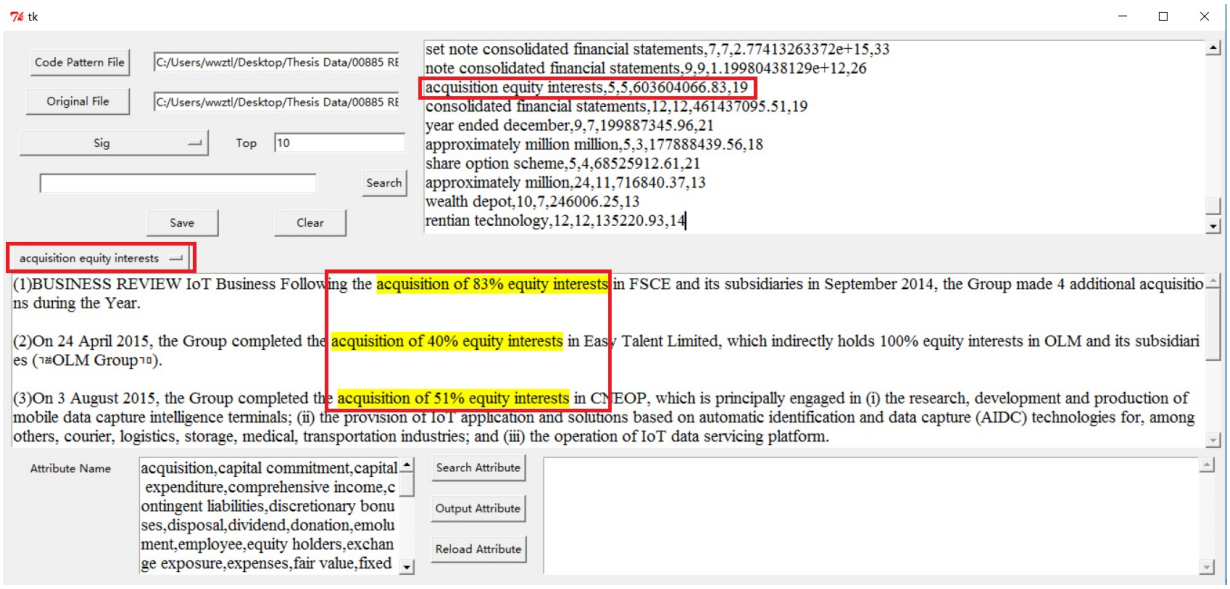


Figure A.11: Part C of Step (6)

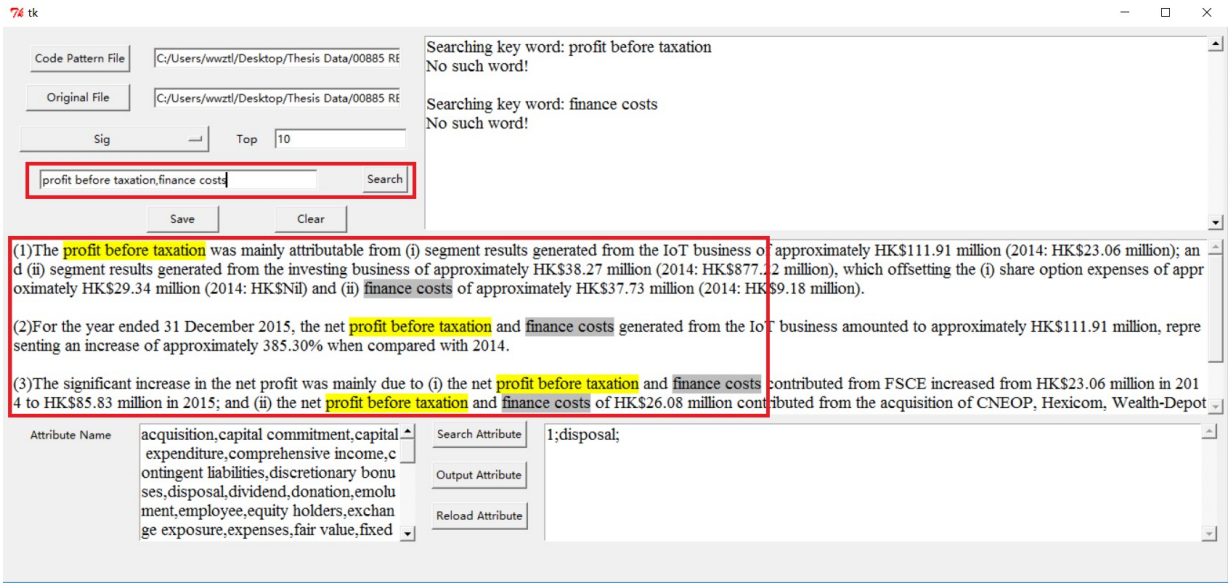


Figure A.12: Find & Locate Context through Domain Knowledge of Users

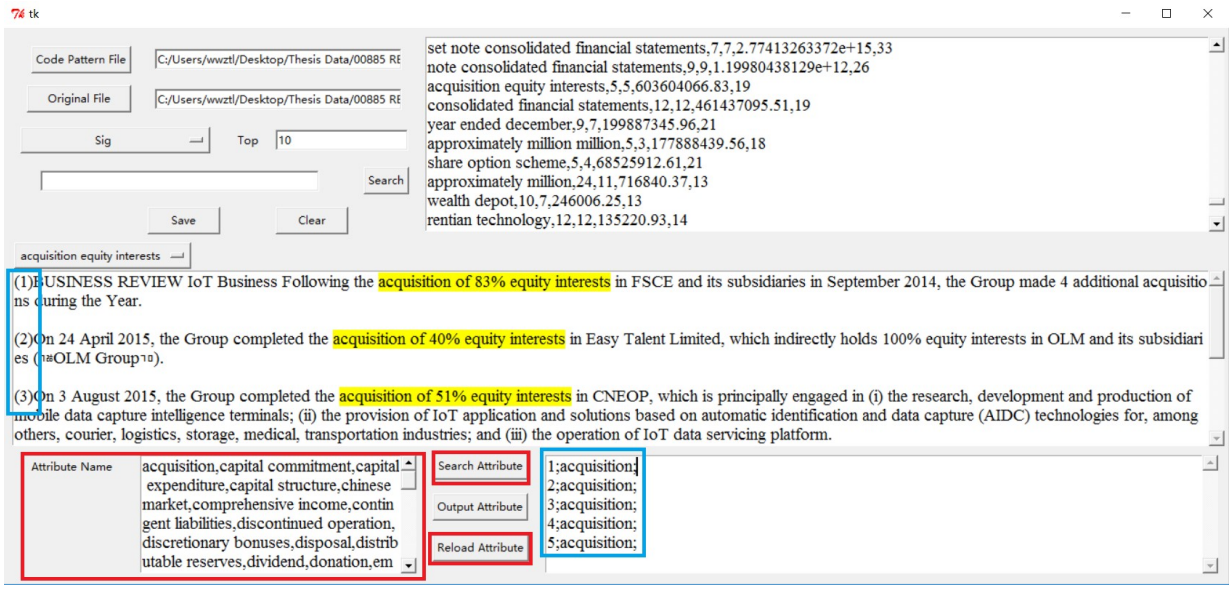


Figure A.13: Use of Local Attribute Name List

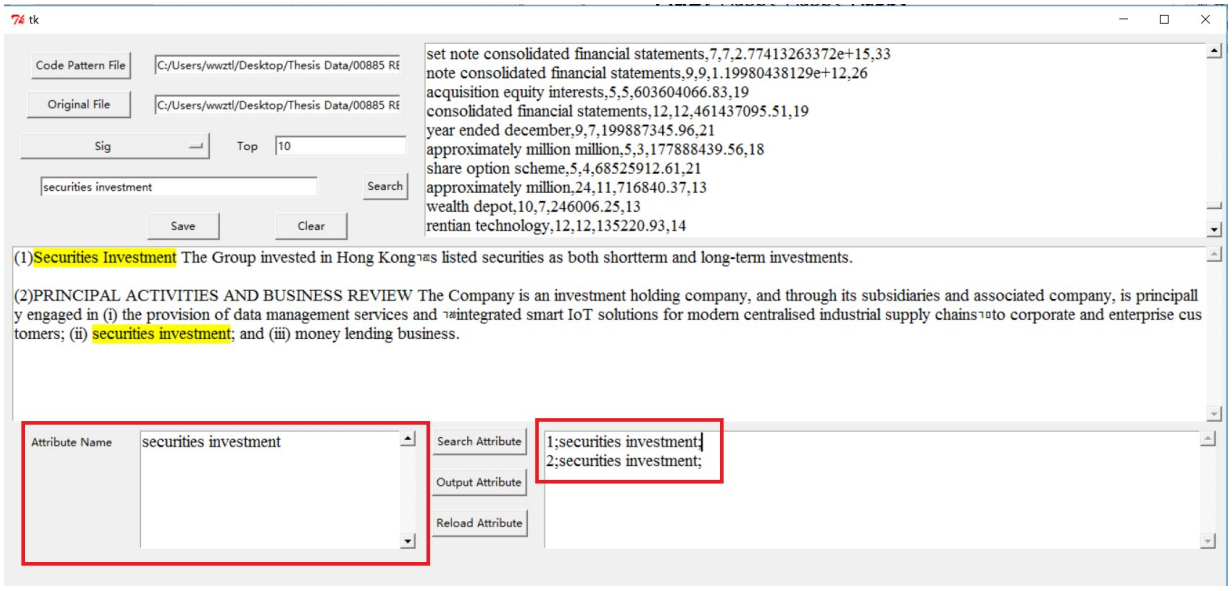


Figure A.14: Use of Attribute Names from Domain Knowledge of the User

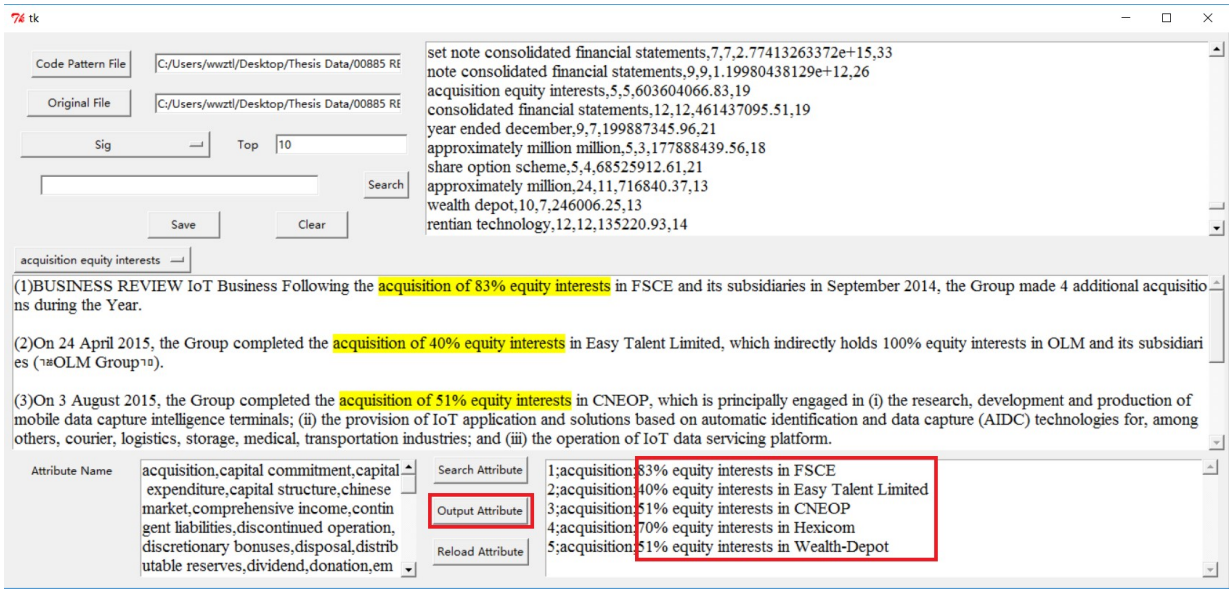
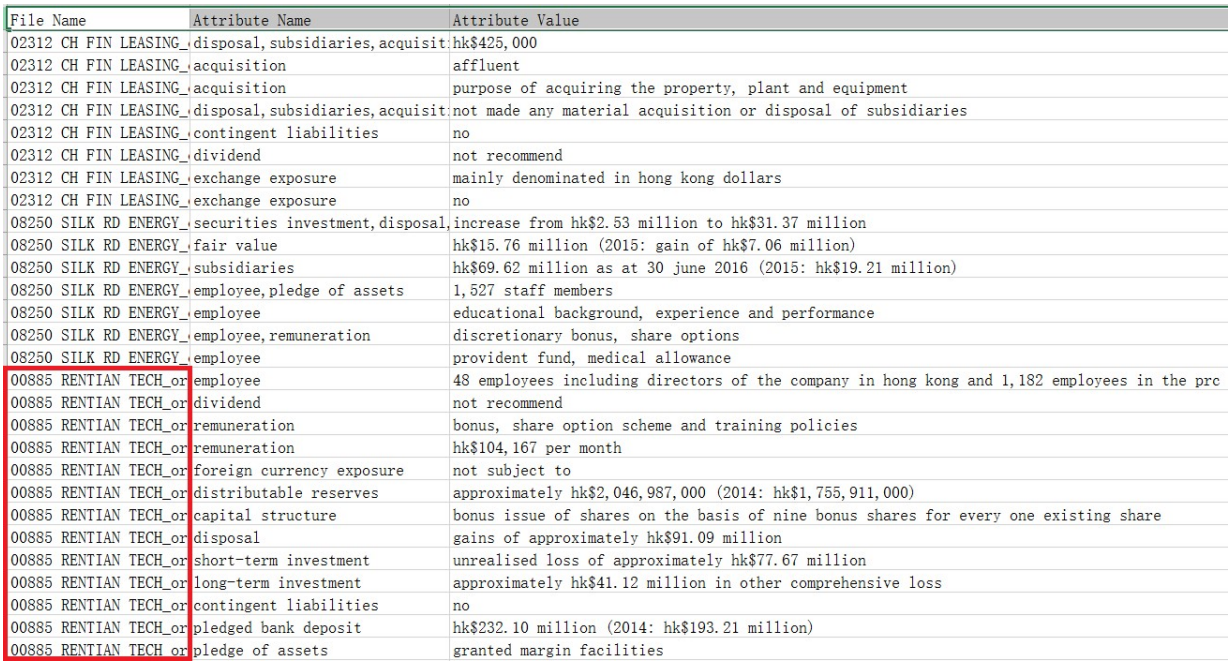


Figure A.15: Attribute Value Validation

Attribute Name	Attribute Value
employee	48 employees including directors of the company in hong kong and 1,182 employees in the prc
dividend	not recommend
remuneration	bonus, share option scheme and training policies
remuneration	hk$104,167 per month
foreign currency exposure	not subject to
distributable reserves	approximately hk$2,046,987,000 (2014: hk$1,755,911,000)
capital structure	bonus issue of shares on the basis of nine bonus shares for every one existing share
disposal	gains of approximately hk$91.09 million
short-term investment	unrealised loss of approximately hk$77.67 million
long-term investment	approximately hk$41.12 million in other comprehensive loss
contingent liabilities	no
pledged bank deposit	hk$232.10 million (2014: hk$193.21 million)
pledge of assets	granted margin facilities
prospect	breakthrough and transformation
market	expand from us$591.7 billion in 2014 to us$1,300.0 billion in 2019 in global
market	optimistic in china
acquisition	83% equity interests in fsce
acquisition	40% equity interests in easy talent limited
acquisition	51% equity interests in cneop
acquisition	70% equity interests in hexicom
acquisition	51% equity interests in wealth-depot
acquisition	5 upstream and downstream iot companies
donation	hk$61,000
redemption	no
plan	gradually realise and improve the synergy within the group
public float	maintained under the listing rules
share option	ending on 5 august 2017, after which no further options will be granted

Figure A.16: Attribute Name and Value Pairs (AVPs) Table for RENTIAN TECH (00885)



File Name	Attribute Name	Attribute Value
02312 CH FIN LEASING_	disposal, subsidiaries, acquisit	hk$425,000
02312 CH FIN LEASING_	acquisition	affluent
02312 CH FIN LEASING_	acquisition	purpose of acquiring the property, plant and equipment
02312 CH FIN LEASING_	disposal, subsidiaries, acquisit	not made any material acquisition or disposal of subsidiaries
02312 CH FIN LEASING_	contingent liabilities	no
02312 CH FIN LEASING_	dividend	not recommend
02312 CH FIN LEASING_	exchange exposure	mainly denominated in hong kong dollars
02312 CH FIN LEASING_	exchange exposure	no
08250 SILK RD ENERGY_	securities investment, disposal,	increase from hk$2.53 million to hk$31.37 million
08250 SILK RD ENERGY_	fair value	hk$15.76 million (2015: gain of hk$7.06 million)
08250 SILK RD ENERGY_	subsidiaries	hk$69.62 million as at 30 june 2016 (2015: hk$19.21 million)
08250 SILK RD ENERGY_	employee, pledge of assets	1,527 staff members
08250 SILK RD ENERGY_	employee	educational background, experience and performance
08250 SILK RD ENERGY_	employee, remuneration	discretionary bonus, share options
08250 SILK RD ENERGY_	employee	provident fund, medical allowance
00885 RENTIAN TECH_or	employee	48 employees including directors of the company in hong kong and 1,182 employees in the prc
00885 RENTIAN TECH_or	dividend	not recommend
00885 RENTIAN TECH_or	remuneration	bonus, share option scheme and training policies
00885 RENTIAN TECH_or	remuneration	hk$104,167 per month
00885 RENTIAN TECH_or	foreign currency exposure	not subject to
00885 RENTIAN TECH_or	distributable reserves	approximately hk$2,046,987,000 (2014: hk$1,755,911,000)
00885 RENTIAN TECH_or	capital structure	bonus issue of shares on the basis of nine bonus shares for every one existing share
00885 RENTIAN TECH_or	disposal	gains of approximately hk$91.09 million
00885 RENTIAN TECH_or	short-term investment	unrealised loss of approximately hk$77.67 million
00885 RENTIAN TECH_or	long-term investment	approximately hk$41.12 million in other comprehensive loss
00885 RENTIAN TECH_or	contingent liabilities	no
00885 RENTIAN TECH_or	pledged bank deposit	hk$232.10 million (2014: hk$193.21 million)
00885 RENTIAN TECH_or	pledge of assets	granted margin facilities

Figure A.17: Local Universal Attribute Name and Value Pairs (AVPs) Database

Appendix B

Applied Numerical Features

Applied numerical features in association with their financial ratios names, symbols, formulas, categories and binarized feature processing are presented in Fig. B.1.

No.	Category	Financial Ratio Name	Formula	Binarized Processing
R1	Liquidity-Solvency	Current Ratio	= Current Assets / Current Liabilities	More than 2, label 1, or label 0
R2	Liquidity-Solvency	Quick Ratio	= (Current Assets – Inventories) / Current Liabilities	More than 1, label 1, or label 0
R3	Liquidity-Solvency	Cash Flows from Operations to Current Liabilities Ratio	= Net Operating Cash Flow / Current Liabilities	More than 0.5, label 1, or label 0
R4	Liquidity-Solvency	Cash Flows from Operations to Debt Ratio	= Net Operating Cash Flow / Total Liabilities	More than 0.25, label 1, or label 0
R5	Liquidity-Solvency	Long-Term Debt to Equity Ratio	= Long-Term Liabilities / Total Shareholders' Equity	Less than 1, label 1, or label 0
R6	Liquidity-Solvency	Debt to Equity Ratio	= Total Liabilities / Total Shareholders' Equity	Less than 2, label 1, or label 0
R7	Capital Structure	Interest Coverage Ratio	= Earnings Before Interest and Tax (EBIT) / Interest Expense	More than 2.5, label 1, or label 0
R8	Capital Structure	Debt to Assets Ratio	= Total Liabilities / Total Assets	Less than 0.85, label 1, or label 0
R9	Capital Structure	Equity Multiplier (EM)	= Total Assets / Total Stockholders' Equity	Less than 2.5, label 1, or label 0
R10	Capital Structure	Debt to Equity Ratio	= Total Liabilities / Total Shareholders' Equity	Less than 1.5, label 1, or label 0
R11	Capital Structure	Current Assets to Total Assets Ratio (CATA)	= Current Assets / Total Assets	Between 0.2 and 0.6, label 1, or label 0
R12	Profitability-Growth	Gross Profit Margin	= (Revenue - Cost of Goods Sold) / Revenue	More than 0.15, label 1, or label 0
R13	Profitability-Growth	Net Profit Margin	= Net Income / Sales Revenue	More than 0.15, label 1, or label 0
R14	Profitability-Growth	Return on Equity (ROE)	= Net Income / Total Stockholders' Equity	More than 0.0955, label 1, or label 0
R15	Profitability-Growth	Return on Invested Capital (ROIC)	= (Net Income - Dividend) / Total Employed Capital	N/A
R16	Profitability-Growth	Operating Profit Margin Ratio	= Operating Income / Total Net Income	N/A
R17	Profitability-Growth	Cash Flows from Operations to Net Income Ratio	= Net Operating Cash Flow / Total Net Income	More than 0.9, label 1, or label 0
R18	Profitability-Growth	Operation Income Growth Rate	= Operation Income in [Year T - Year (T-1)] / Year (T-1)	More than 0, label 1, or label 0
R19	Operation Capacity	Inventory Turnover Rate	= Sales Revenue / Average Inventory	More than 3, label 1, or label 0
R20	Operation Capacity	Current Assets Turnover Rate	= Sales Revenue / Total Current Assets	More than 1, label 1, or label 0
R21	Operation Capacity	Total Assets Turnover Rate	= Sales Revenue / Total Assets	More than 0.8, label 1, or label 0
R22	Investment Value	Price-Earnings Ratio	= Price per Share / Earnings per Share	Less than 20, label 1, or label 0
R23	Investment Value	Price-Net Book Value Ratio	= Market Price per Share / Net Book Value per Share	Less than 2.5, label 1, or label 0
R24	Investment Value	Price-Sales Ratio	= Market Capitalization / Total Revenue	Less than 1.5, label 1, or label 0
R25	Investment Value	Dividend Yield	= Annual Dividend per Share / Price per Share	More than 0.0415, label 1, or label 0

Figure B.1: Applied Numerical Features

Appendix C

Applied c-TP2K Textual Features

Applied c-TP2K textual features in association with their names, symbols, representative corresponding important attribute names and binarized feature processing are presented in Fig. C.1.

Numeber	Textual Feature Name	Representative Correspodning Attribute Names	Binarized Processing
T1	Risks / Uncentaties	e.g. contingent liabilities / discountinued operations	Subjective judgement by user, Positive (1) or Negative (0)
T2	Prospects / Outlooks	e.g. public float / acquisition / fund raising	Subjective judgement by user, Positive (1) or Negative (0)
T3	Remunaration / Emolument	e.g. discretionary bonuses / house allowance	Yes (1) or Not (0)
T4	Exchange Expousre	e.g. currency risk / hedging tool	For Risk, Yes (0) or Not (1); For Tool, Yes (1) or Not (0)
T5	Employee / Human Resource	employee	Compare with last year, more (1) or less (0)
T6	Dividend	dividend	Yes (1) or Not (0)
T7	Donation	donation	N/A
T8	Incentive Policy	e.g. share option scheme	Yes (1) or Not (0)
T9	Pledge of Assets / Deposit	e.g. pledge of assets / pledge of bank deposit	Yes (0) or Not (1)
T10	Major Customers / Suppliers	e.g. major customer / major supplier	N/A

Figure C.1: Applied c-TP2K textual features

Appendix D

Definition of Evaluation Metrics

Overall Accuracy, F-measure, Type I Error Rate and ROC Area are common evaluation metrics in bankruptcy problems, in accordance with the current literature [11, 58, 16, 68, 42, 34, 47, 41]. As it is the first time to propose 3-class classification in bankruptcy problems, we hereby provide a detailed definition of the evaluation metrics as follows.

In this study, for a 3-class (A, B and C) classification system, the confusion matrix is depicted in Fig. D.1, where the matrix entries $e11$, $e12$, $e13$, $e21$, $e22$, $e23$, $e31$, $e32$, $e33$ are all integers.

	Predicted as "A"	**Predicted as "B"**	**Predicted as "C"**
Actual as "A"	e11	e12	e13
Actual as "B"	e21	e22	e23
Actual as "C"	e31	e32	e33

Figure D.1: Confusion Matrix of 3-Class Classification System

The Overall Accuracy (OA) in percentage (%) is defined as:

$$OverallAccuracy = \frac{e11 + e22 + e33}{e11 + e12 + e13 + e21 + e22 + e23 + e31 + e32 + e33} * 100 \quad \text{(D.1)}$$

The actual number of class A, class B and class C samples are defined respectively as:

$$Actual_A = e11 + e12 + e13 \tag{D.2}$$

$$Actual_B = e21 + e22 + e23 \tag{D.3}$$

$$Actual_C = e31 + e32 + e33 \tag{D.4}$$

The number of samples predicted as class A, class B and class C are defined as:

$$Predicted_A = e11 + e21 + e31 \tag{D.5}$$

$$Predicted_B = e12 + e22 + e32 \tag{D.6}$$

$$Predicted_C = e13 + e23 + e33 \tag{D.7}$$

In this study, if class A is defined as positive class, then class B and class C are defined as negative classes. Thus, $Precision_A$, $Recall_A$ and $Fmeasure_A$ are defined as follows.

$$Precision_A = \frac{e11}{Predicted_A} \tag{D.8}$$

$$Recall_A = \frac{e11}{Actual_A} \tag{D.9}$$

$$Fmeasure_A = \frac{2 * Precision_A * Recall_A}{Precision_A + Recall_A} \tag{D.10}$$

In this study, if class B is defined as positive class, then class A and class C are defined as negative classes. Thus, $Precision_B$, $Recall_B$ and $Fmeasure_B$ are defined as follows.

$$Precision_B = \frac{e22}{Predicted_B} \tag{D.11}$$

$$Recall_B = \frac{e22}{Actual_B} \tag{D.12}$$

$$Fmeasure_B = \frac{2 * Precision_B * Recall_B}{Precision_B + Recall_B} \tag{D.13}$$

In this study, if class C is defined as positive class, then class A and class B are defined as negative classes. Thus, $Precision_C$, $Recall_C$ and $Fmeasure_C$ are defined as follows.

$$Precision_C = \frac{e33}{Predicted_C} \tag{D.14}$$

$$Recall_C = \frac{e33}{Actual_C} \tag{D.15}$$

$$Fmeasure_C = \frac{2 * Precision_C * Recall_C}{Precision_C + Recall_C} \tag{D.16}$$

In this study, $Fmeasure$ is defined as the weighted average of $Fmeasure_A$, $Fmeasure_B$ and $Fmeasure_C$, as follows.

$$Fmeasure = \frac{Fmeasure_A * Actual_A + Fmeasure_B * Actual_B + Fmeasure_C * Actual_C}{Actual_A + Actual_B + Actual_C} \tag{D.17}$$

In this study, False Positive Rate (FPR) is defined as the number of false positives divided by the number of negatives.

If class A is defined as positive, then class B and class C are defined as negative.

Then FPR_A is defined as follows:

$$FPR_A = \frac{e21 + e31}{Actual_B + Actual_C} \tag{D.18}$$

If class B is defined as positive, then class A and class C are defined as negative.

Then FPR_B is defined as follows:

$$FPR_B = \frac{e12 + e32}{Actual_A + Actual_C} \tag{D.19}$$

If class C is defined as positive, then class A and class B are defined as negative.

Then FPR_C is defined as follows:

$$FPR_C = \frac{e13 + e23}{Actual_A + Actual_B} \tag{D.20}$$

In this study, Type I Error (Rate) is defined as the weighted average of FPR_A, FPR_B and FPR_C, as follows:

$$TypeIError = \frac{FPR_A * Actual_A + FPR_B * Actual_B + FPR_C * Actual_C}{Actual_A + Actual_B + Actual_C} \tag{D.21}$$

For a binary classification system, a ROC curve is obtained by plotting the True Positive Rate, i.e. Recall, against False Positive Rate, i.e. Type I Error, via varying the prediction threshold of the classifier. Hence, the ROC Area is obtained by computing the area under the ROC curve.

Hence, in this study, for a 3-class classification system: $ROCArea_A$ is obtained by plotting $Recall_A$ against FPR_A; $ROCArea_B$ is obtained by plotting $Recall_B$ against FPR_B; $ROCArea_C$ is obtained by plotting $Recall_C$ against FPR_C. All these computations are obtained via the software WEKA [28].

Finally, in this study, $ROCArea$ is defined as the weighted average of $ROCArea_A$, $ROCArea_B$ and $ROCArea_C$, as follows.

$$ROCArea = \frac{ROCArea_A * Actual_A + ROCArea_B * Actual_B + ROCArea_C * Actual_C}{Actual_A + Actual_B + Actual_C} \tag{D.22}$$