

4-2015

Extracting Implicit Knowledge and Inter-relationships from Initial Public Offering (IPO) Prospectus for Pricing Prediction

Jie Tao

Follow this and additional works at: <https://scholar.dsu.edu/theses>

Recommended Citation

Tao, Jie, "Extracting Implicit Knowledge and Inter-relationships from Initial Public Offering (IPO) Prospectus for Pricing Prediction" (2015). *Masters Theses & Doctoral Dissertations*. 408.
<https://scholar.dsu.edu/theses/408>

This Dissertation is brought to you for free and open access by Beadle Scholar. It has been accepted for inclusion in Masters Theses & Doctoral Dissertations by an authorized administrator of Beadle Scholar. For more information, please contact repository@dsu.edu.

Extracting Implicit Knowledge and Inter-relationships from Initial Public Offering (IPO) Prospectus for Pricing Prediction

A dissertation submitted to Dakota State University in partial fulfillment of the
requirement for the degree of

Doctor of Science

in

Information Systems

17 April 2015

By

Jie Tao

KARL E. MUNDT LIBRARY
Dakota State University
Madison, SD 57042-1799

Dissertation Committee:

Dr. Amit V. Deokar (Co-Chair)

The Pennsylvania State University – Erie

Dr. Yen-Ling Chang (Co-Chair)

Dakota State University

Dr. John Nelson

Dakota State University



DISSERTATION APPROVAL FORM

This dissertation is approved as a credible and independent investigation by a candidate for the Doctor of Science in Information Systems degree and is acceptable for meeting the dissertation requirements for this degree. Acceptance of this dissertation does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department or university.

Student Name: Jie Tao

Dissertation Title: Extracting Implicit Knowledge and Inter-relationships from Initial Public Offering (IPO) Prospectus for Pricing Prediction

A handwritten signature in black ink, appearing to read "Amit Deokar", is written over a horizontal line.

Dissertation Chair/Co-Chair: Dr. Amit Deokar

Date: April 10, 2015

A handwritten signature in black ink, appearing to read "Yen-Lin Chang", is written over a horizontal line.

Dissertation Chair/Co-Chair: Dr. Yen-Lin Chang

Date: April 10, 2015

A handwritten signature in black ink, appearing to read "John Nelson", is written over a horizontal line.

Committee member: Dr. John Nelson

Date: April 10, 2015

Acknowledgment

I would like to express the deepest and most sincere appreciation toward my dissertation committee co-chairs, Dr. Amit Deokar and Dr. Yenling Chang, for their persistent supportive attitude and effort, as well as the substance of geniuses. They helped me initiate the topic of this dissertation topic, and all along with the adventure in pursuing the goal of this project. Without their guidance and help, the achievement of this dissertation would not have been possible.

I would also like to thank my dissertation committee member, Dr. John Nelson. Dr. Nelson has provided constructive comments, toward both the content and the presentation of the dissertation. By addressing these comments, the quality of this dissertation is assured.

In addition, I would like to thank other faculty members and staffs at Dakota State University, who helped substantially in this dissertation project. I would also like to express the appreciation to my parents — without their support all the time, I would never have finished this project.

Abstract

Initial Public Offering (IPO) process and the associated pricing strategies are of much interest to researchers and practitioners (e.g., underwriters, investors) in the finance and accounting domains. IPO prospectuses, regulated by Security Exchange Committee (SEC), serve as the most reliable publicly available information source in the IPO process. IPO prospectuses disclose a variety of information; however, traditional studies do not leverage the rich knowledge hidden in the vast textual information within them. The research gap can be partially attributed to the lack of an underlying formal knowledge structure to support the extraction of the implicit knowledge from the prospectuses, as well as to the absence of quantitative metrics that reflect management's outlook and awareness embedded in the prospectuses. The primary research question addressed in this work is: *"How do the management's awareness of risks (expressed via the emphasized mentions in the Risk Factors sections of the Form 424 filings), and confidence about the firm's outlook (expressed through the sentiments in the MD&A sections) affect IPO valuations?"* The major research problem could be further broken down into two research goals: a) to develop an actionable knowledge structure for guiding the extraction, storing the results, and facilitate reasoning of the knowledge hidden in the textual content of the IPO prospectuses; and b) to utilize the knowledge structure developed above, as well as the predictive models, to estimate pricing volatility prior to and right after the IPO date. In order to identify and quantify such inter-relationships, an underlying knowledge structure needs to be constructed and updated with minimal manual interventions for efficient knowledge acquisition and accurate knowledge representation purposes. In this dissertation, to bridge aforementioned research gaps, I proposed a text analytics framework for assisting the investment and underwriting decision making processes. Two major components existing in the proposed framework, namely the ontology enrichment methodology that updates the ontology in real time and online mode, and the predictive modeling techniques using the extracted information based on the ontology for predicting IPO pricing. The proposed framework is then developed in the form of a research prototype, which is used to predict pricing trends

during and after the IPO process. I use real world data to evaluate the framework itself as well as the prediction results through a set of experiments, which yield promising results.

Design science research methodology is applied as the methodological framework in this study. Two motivational scenarios are provided to illustrate the significance and relevance of this study. The searching and developing process of a solution is documented in detail. I have compared our approach to the existing body of research and illustrated its novelty. Further, I have evaluated the proposed IT artifacts (the analytical framework), first through feasibility and functionality testing and second through an experimental approach for analyzing efficiency and accuracy.

The proposed analytical framework is evaluated by various means. First, a case study is designed to evaluate the functionalities and efficiencies of the framework. Second, the practical relevance of the framework is evaluated through the results of the predictive models. Third, the design artifacts are also evaluated against the design requirements drawn from existing literature. The evaluation results in this study are satisfying, which indicate the promising prospects of this project in practice.

There are two key research contributions of this work: a) an (semi-) automatic approach for enriching the specifications of domain knowledge bases (i.e. ontologies) is developed and evaluated, as an underlying knowledge structure for the analytical process. The approach is unique in the sense of incorporating feature-based word sense disambiguation and relation extraction methods in the process; b) several predictive models are designed based on extracted knowledge from the prospectus, for the purpose of predicting pre- and post-IPO pricing volatility. The results of this phase of the study ensure its practical relevance. In addition to these two primary contributions, two metrics are also designed, as a proxy of the management's awareness of risks and management's confidence regarding the organization's future operations. This metrics are based on the textual contents in the more informative sections in the prospectuses (i.e. Risk Factors, Management Discussions and Analysis) and to the best of our knowledge, these metrics are the first of its kind to quantify such information. Further, the analytical framework and development approaches of the design artifacts can be adopted in other application domains such as healthcare informatics, social media analysis, and so forth.

Declaration

I hereby certify that this dissertation constitutes my own product, that where the language of others is set forth, quotation marks so indicate, and that appropriate credit is given where I have used the language, ideas, expressions or writings of another.

I declare that the dissertation describes original work that has not previously been presented for the award of any other degree of any institution.

Signed,

A handwritten signature in black ink, appearing to read 'Jie Tao', is written over a horizontal line.

Jie Tao

TABLE OF CONTENTS

Acknowledgment	i
Abstract	iii
Declaration	v
TABLE OF CONTENTS	vi
LIST OF TABLES	x
LIST OF FIGURES	xii
CHAPTER 1	1
INTRODUCTION	1
1.1 Background	1
1.2 Problem Statement	2
1.2.1 Motivations of the Study	3
1.2.2 Research Problem	4
1.3 Research Objectives	6
1.4 Scope and Contributions of the Study	7
1.5 Structure of the Dissertation	11
CHAPTER 2	12
LITERATURE REVIEW	12
2.1 IPO Pricing Determinants: Theoretical Foundation	12
2.2 IPO Valuation and Informative Contents of Prospectuses	14
2.3 Ontology Enrichment from Textual Corpus	16
2.4 Design Requirements	17
CHAPTER 3	20
RESEACH METHODOLOGY	20
3.1 Design Science Research Methodology	20

3.1.1 Design as an Artifact.....	20
3.1.2 Problem Relevance	21
3.1.3 Design Evaluation	21
3.1.4 Research Contributions	22
3.1.5 Research Rigor.....	22
3.1.6 Design as a Search Process	22
3.1.7 Communication of Research.....	23
CHAPTER 4	24
THEORY AND ARTIFACT DESIGN	24
4.1 An Overview of the Analytical Framework.....	24
4.1.1. Information Extraction Module	26
4.1.2. Ontology Enrichment Module	26
4.1.3 Predictive Analytics Module.....	27
4.1.4 IPO Ontology	27
4.2 Information Extraction Module	29
4.3 Ontology Enrichment Module	31
4.3.1 Domain Specific Term Extraction and Selection.....	32
4.3.2 Word Sense Disambiguation.....	34
4.3.3 Ontology Integration.....	36
4.3.4. Semantic Relation Extraction	39
4.4. Predictive Analytics Module.....	43
4.4.1 Model Building	44
4.4.2. Modeling Techniques.....	47
4.4.3 Model Validation and Selection Strategy	49
4.4.4. Result Validation and Communication.....	52

CHAPTER 5	53
DEMONSTRATION AND EVALUATION	53
5.1 Ontology Enrichment Module	53
5.1.1. Term Extraction Component.....	54
5.1.2 Term Selection Component	55
5.1.3 Ontology Integration Component	56
5.1.4 Relation Extraction Component.....	57
5.1.5 Evaluation of the Ontology Enrichment Module: A Preliminary Case Study.	58
5.2 Predictive Modeling Module	64
5.3 Study Data.....	67
5.3.1 Descriptive Statistics.....	68
5.4 Experiments, Results and Discussions.....	71
5.4.1. Alternative 1: Using Risk Features and Predictors from Prior Studies	72
5.4.2. Alternative 2: Using MD&A Features and Predictors from Prior Studies	80
5.4.3. Alternative 3: Using Both Risk and MD&A Features and Predictors from Prior Studies.....	88
5.4.4. Alternative 4: Using Aggregated Predictors and Predictors from Prior Studies	95
5.4.5 Implications to Practitioners	103
CHAPTER 6	105
CONCLUSIONS.....	105
6.1 Concluding Remarks.....	105
6.2 Contributions Revisited	107
6.3 Lessons Learned.....	108
6.3.1 Method-oriented Issues	108
6.3.2 Application-oriented Issues	110

6.4 Limitations and Future Steps	112
REFERENCES	114
Appendix A. Research Plan: Data Analytics Lifecycle	123

LIST OF TABLES

Table 4.1. Noun Phrase Patterns	33
Table 4.2. Selected lexico-syntactic patterns and examples	40
Table 4.3. Predictors Used in Predictive Models.....	46
Table 4.4. Modeling Techniques Used in the Predictive Modeling Module	49
Table 5.1. Rules Used in Deep Cleansing.....	56
Table 5.2. Summary Statistics of All Variables.....	68
Table 5.3(a). Confusion Matrices of PRCREV in Alternative 1	73
Table 5.3(b). Accuracy Metrics of PRCREV in Alternative 1	73
Table 5.3(c) Efficiency Metrics of PRCREV in Alternative 1	74
Table 5.4(a). Confusion Matrices of X1stDay in Alternative 1	76
Table 5.4(b). Accuracy Metrics of X1stDay in Alternative 1.....	77
Table 5.4(c) Efficiency Metrics of X1stDay in Alternative 1	77
Table 5.5(a). Confusion Matrices of PRCREV in Alternative 2	80
Table 5.5(b). Accuracy Metrics of PRCREV in Alternative 2	80
Table 5.5(c) Efficiency Metrics of PRCREV in Alternative 2	81
Table 5.6(a). Confusion Matrices of X1stDay in Alternative 2.....	84
Table 5.6(b). Accuracy Metrics of X1stDay in Alternative 2.....	84
Table 5.6(c) Efficiency Metrics of X1stDay in Alternative 2	85
Table 5.7(a). Confusion Matrices of PRCREV in Alternative 3	88
Table 5.7(b). Accuracy Metrics of PRCREV in Alternative 3	89
Table 5.7(c) Efficiency Metrics of PRCREV in Alternative 3	89
Table 5.8(a). Confusion Matrices of X1stDay in Alternative 3.....	92
Table 5.8(b). Accuracy Metrics of X1stDay in Alternative 3.....	92
Table 5.8(c) Efficiency Metrics of X1stDay in Alternative 3	93
Table 5.9(a). Confusion Matrices of PRCREV in Alternative 4	96
Table 5.9(b). Accuracy Metrics of PRCREV in Alternative 4	96
Table 5.9(c) Efficiency Metrics of PRCREV in Alternative 4	97
Table 5.10(a). Confusion Matrices of X1stDay in Alternative 4.....	99

Table 5.10(b). Accuracy Metrics of X1stDay in Alternative 4.....	100
Table 5.10(c) Efficiency Metrics of X1stDay in Alternative 4	100

LIST OF FIGURES

Figure 4.1. The Architecture of the Proposed Analytical Framework.....	25
Figure 4.2 Overview of IPO Ontology	28
Figure 4.3. Segmentation of IPO Prospectus	30
Figure 4.4 Semantic Annotations and Ontology Population/Instantiation	31
Figure 4.5. The F-WSD Algorithm.....	36
Figure 4.6. <i>WuP</i> Calculation Example.....	38
Figure 4.7. Flow chart of relation extraction	41
Figure 5.1. Ontology Enrichment Pipeline	54
Figure 5.2. Grammar Tree and Clause Structure of the Example Sentence	58
Figure 5.3. A Snippet of <i>IPO-Ontology</i>	60
Figure 5.4. Efficiency Comparison Between Different Methods	61
Figure 5.5. Coverage/Correctness Comparison Between Different Methods.....	62
Figure 5.6. Quality Metrics of IPO-Extractor on Select Text Corpus	63
Figure 5.7. Quality Comparison Between Different Approaches.....	64
Figure 5.8. FOCAS-IE Analytical Pipeline	65
Figure 5.9. AUC Diagram of Predictive Models toward PRCREV in Alternative 1	74
Figure 5.10(a) Lift Curve with Training Data Set toward PRCREV in Alternative 1.....	75
Figure 5.10(b) Lift Curve with Validation Data Set toward PRCREV in Alternative 1 ..	75
Figure 5.10(c) Lift Curve with Testing Data Set toward PRCREV in Alternative 1	76
Figure 5.11. AUC Diagram of Predictive Models toward X1stDay in Alternative 1.....	78
Figure 5.12(a) Lift Curve with Training Data Set toward X1stDay in Alternative 1	79
Figure 5.12(b) Lift Curve with Validation Data Set toward X1stDay in Alternative 1....	79
Figure 5.12(c) Lift Curve with Testing Data Set toward X1stDay in Alternative 1.....	79
Figure 5.13. AUC Diagram of Predictive Models toward PRCREV in Alternative 2	82
Figure 5.14(a) Lift Curve with Training Data Set toward PRCREV in Alternative 2.....	83
Figure 5.14(b) Lift Curve with Validation Data Set toward PRCREV in Alternative 2 ..	83
Figure 5.14(c) Lift Curve with Testing Data Set toward PRCREV in Alternative 2	83
Figure 5.15. AUC Diagram of Predictive Models toward X1stDay in Alternative 2.....	86

Figure 5.16(a) Lift Curve with Training Data Set toward X1stDay in Alternative 2	87
Figure 5.16(b) Lift Curve with Validation Data Set toward X1stDay in Alternative 2....	87
Figure 5.16(c) Lift Curve with Testing Data Set toward X1stDay in Alternative 2	87
Figure 5.17. AUC Diagram of Predictive Models toward PRCREV in Alternative 3	90
Figure 5.18(a) Lift Curve with Training Data Set toward PRCREV in Alternative 3	91
Figure 5.18(b) Lift Curve with Validation Data Set toward PRCREV in Alternative 3 ..	91
Figure 5.18(c) Lift Curve with Testing Data Set toward PRCREV in Alternative 3	91
Figure 5.19. AUC Diagram of Predictive Models toward X1stDay in Alternative 3	94
Figure 5.20(a) Lift Curve with Training Data Set toward X1stDay in Alternative 3	94
Figure 5.20(b) Lift Curve with Validation Data Set toward X1stDay in Alternative 3....	94
Figure 5.20(c) Lift Curve with Testing Data Set toward X1stDay in Alternative 3	95
Figure 5.21. AUC Diagram of Predictive Models toward PRCREV in Alternative 4	97
Figure 5.22(a) Lift Curve with Training Data Set toward PRCREV in Alternative 4	98
Figure 5.22(b) Lift Curve with Validation Data Set toward PRCREV in Alternative 4 ..	98
Figure 5.22(c) Lift Curve with Testing Data Set toward PRCREV in Alternative 3	99
Figure 5.23. AUC Diagram of Predictive Models toward X1stDay in Alternative 4	101
Figure 5.24(a) Lift Curve with Training Data Set toward X1stDay in Alternative 4	102
Figure 5.24(b) Lift Curve with Validation Data Set toward X1stDay in Alternative 4..	102
Figure 5.24(c) Lift Curve with Testing Data Set toward X1stDay in Alternative 4	102
Figure 6.1. “Delta Analysis” on IPO Prospectuses	112
Figure A.1 DALC used in This Project as Research Plan	125

CHAPTER 1

INTRODUCTION

1.1 Background

Initial Public Offering (IPO) refers to the process that an organization determines to raise capital from the public market for support a business decision (Jain & Kini, 1999). Understanding and predicting various phenomena in IPO processes have become one of the most attractive puzzles in the finance domain (Babich & Sobel, 2004; Lowry & Schwert, 2004; Varshney & Robinson, 2004). *Prospectus* refers to the documents containing certain specific contents, disclosing relevant information, regulated by the *Security Exchange Committee* (SEC) before any new security can be issued to the public market (Bhabra & Pettway, 2003). Vast amounts of textual information are available in IPO prospectus. Among different sections of the prospectus, Hanley and Hoberg (2010) mention the following four sections as generally more informative than the others – *Prospectus Summary*, *Risk Factors*, *Use of Proceeds*, and *Management's Discussion and Analysis (MD&A)*. Based on a close scrutiny of sample prospectuses, as well as review of the financial literature, it is found that two sections in the prospectus, *Risk Factors* and *MD&A*, to contain richer textual information than others. Studies have been conducted in order to reveal the inter-relationships between the information disclosed in the prospectus and the phenomena in the IPO process; however, most studies have only focused on the quantitative contents or the relevant keywords in the prospectus documents (Daily, Certo, Dalton, & Roengpitya, 2003; Rajan & Servaes, 1997; Ritter, 1991). Thus, the rich implicit knowledge, such as the implicit semantics embedded in the textual contents, has not been harnessed. Recently, researchers have started to recognize the importance of this issue (Hanley & Hoberg, 2012). Thus, there is an increasing demand of extracting relevant domain-specific knowledge hidden in the textual parts of the prospectus for the

analytical purposes. Further, given that an issuing firm typically releases multiple versions of its prospectus, such text mining exercise needs to be conducted in an iterative fashion on various versions. Some pilot studies have tapped into the textual contents of the prospectus (Arnold, Fische, & North, 2010; S. P. Ferris, Hao, & Liao, 2012). Yet such studies concentrate only on identifying keywords (i.e. *risks*, *finance*, *loss*, etc.) in a given section (*Risk Factors*, etc.). Due to the lack of a comprehensive and formalized text analytics framework that leverages domain knowledge, significant information such as the semantic relations between the keywords (concepts) still remains unrevealed.

Particularly within the IPO process, understanding the determinants and dynamics behind the pricing strategies has been of keen interest to researchers and practitioners in different business domains. Although there is a large body of work attempting to understand the “underpricing” phenomenon – such as the work by Ferris et al. (2013), Hanley and Hoberg (2010), Loughran and McDonald (2013), Loughran and Ritter (2004) – however, much less work has been done to explain the sentiments of the issuer/underwriters as it relates to the IPO price revisions. Most studies in this area focus on analyzing the post-IPO market/investor sentiments and their effects on the demand of IPOs, via sentiment analysis of post issuance financial news, reports, and social media (Geva & Zahavi, 2014; M.-C. Lin, Lee, Kao, & Chen, 2011). Open research issues remain regarding how the management’s confidence about the offering as well as the company’s future performance outlook, as well as its awareness of the risks, affects IPO price volatilities. Researchers have argued that the pre-IPO pricing is of equal importance, if not more, compared to the post-IPO pricing, and that the key to understanding the pre-IPO pricing is through sentiment analysis of management’s confidence about the company’s future outlook and awareness of risks (Cornelli, Goldreich, & Ljungqvist, 2006).

1.2 Problem Statement

In this section, I further elicit the research problem along two dimensions. Following subsections presents the research problem in a more detailed manner: Section

1.2.1 provides the motivation of the study via two scenarios; while Section 1.2.2 presents the research problem in the form of two research questions.

1.2.1 Motivations of the Study

Two typical scenarios are presented below for illustrating the motivations of this study. These scenarios are selected from the perspectives of the (potential) *investors* and the *issuer/underwriters* of an IPO process. The (potential) investors refer to the ultimate buyers of the stock in the open market; the issuer is the firm that releasing the stock to the public; while the underwriter is the agency that representing the issuer in the IPO process. The prospectus serves as the most credible and informative source for the issuer/underwriter providing information regarding the firm to the potential investors. Following discussions provide motivations of the study from both the investor's and the underwriters' perspectives – which are two major involved parties in the IPO process.

1. "Underpricing" and "overpricing" are two common phenomena in the IPO process.

For instance, Facebook Inc. is a typical example of *overpricing* – the stock price started descending after the initial offering, and continued to decline almost 50% over the first several months. On the other hand, "underpricing" is more intentionally set by the underwriters, for avoiding risks and obtaining higher returns. When a version of the prospectus is released, the market (comprised of potential investors, either institutional and individual) reacts to it by speculating the price of the final offer based on the information disclosed in the documents. However, the average investors' lacking of financial expertise and domain knowledge for retrieving and comprehending such implicit yet crucial information leads to inaccurate estimations of the stock price trends – this is particularly common with individual investors. Thus, they need a decision support system to assist their investment decision-making process, by analyzing the informative contents in the IPO prospectuses. It has also been emphasized in the literature that analyzing informative contents within the prospectus would be beneficial in estimating the *true* price of a stock (Hanley & Hoberg, 2010).

2. On the other hand, the underwriters representing the issuing firms also need a normalized knowledge structure as a *reference* for composing the prospectus

documents. Underwriters play a very significant role in the IPO process (Ferris et al., 2012). Despite the attitudes and writing patterns of different underwriters, a prospectus of good quality requires certain expertise and insights within the domain. For example, compared to a veteran underwriter, a novice underwriter may lack of adequate skills and experiences, which could possibly lead to an unsuccessful IPO. Furthermore, underwriters need a reference knowledge framework, as the core of a decision support system, to support the decision making process within underwriting and associated book-building process (i.e. be aware of the consequences regarding the disclosure of certain types of information – *risks*, *management's prospect*, etc.). A *formalized, comprehensive, and industry-independent* knowledge structure would assist him/her when composing a prospectus document.

To summarize, the design artifacts in this study will help reading (from the investors' perspective) and support the investment decision making process; as well as writing the prospectus (from the underwriters' perspective) and assist the composition of prospectuses. Such motivations lead to the concrete research questions of this study.

1.2.2 Research Problem

The main research problem of this project can be further broken down into two research questions (RQ1 and RQ2), which are stated as follows:

RQ1. *What is an actionable underlying knowledge structure for guiding the extraction, enabling the storage, and facilitating reasoning of knowledge hidden in the textual contents of IPO prospectuses?*

As mentioned in the motivational scenarios, a formal knowledge structure, as the core of the decision support system, is very important for both investors and issuers/underwriters. As a specification of the underlying knowledge base, a domain ontology containing the terms and relations regarding the IPO process needs to be created. However, manually constructing and enriching ontologies are demanding and tedious activities (Zhou, 2007). In order to overcome such shortcomings, automated

learning techniques, as a gateway to (semi-) automatically create or extend ontologies using machine learning techniques, have been well studied. Information Extraction (IE), as a sub-field of text analytics, is an effective approach for extracting implicit knowledge from unstructured sources (i.e. textual data, etc.) (Cowie & Wilks, 2000). Grishman (1997) defined IE as “the identification of instances of a particular class of events or relationships in a natural language text, and the extraction of the relevant arguments of the event or relationship.” Thus, IE has the potential to serve as the knowledge-discovering technique for analyzing the textual content in the IPO prospectuses. IE has been proven to be one of the most important techniques in terms of enriching ontologies. Among the various sources for IE, text mining turns out to be one of the most effective approaches (Wong, Liu, & Bennamoun, 2012). An important application of text mining in that regard is to extract and disambiguate relevant terms from a domain-specific corpus, as well as to identify relations, for the purpose of enriching knowledge bases – and ultimately support the knowledge acquisition and subsequent activities in the knowledge management process. The proposed ontology learning approach should be designed and evaluated to minimize human intervention in the ontology enrichment process, while yielding quality results.

***RQ2.** Which predictive model(s) is better suited to estimate price volatility prior to and right after the IPO date, based on the knowledge extracted from the IPO prospectus based on the ontology created above?*

RQ2 should be interpreted as following: “How do the management’s awareness of risks (expressed via the emphasized mentions in the Risk Factors sections of the Form 424 filings, and confidence about the firm’s outlook (expressed through the sentiments in the MD&A sections) affects IPO valuations?” This research question relates to the management’s prospects toward the firm’s issuance and future performances and its awareness of the risks the firm would encounter currently and in the future, as well as their impacts on the pre-IPO price adjustment(s) as well as post-IPO initial returns. As discussed in the finance literature (Ferris et al., 2013; Hanley & Hoberg, 2010; Loughran & McDonald, 2013), during the IPO process starting from the disclosure of the initial

range of the offering price in the initial prospectus (Form S-1) to the offering price is determined in the final prospectus (Form 424), there might be none, one, or more price adjustments. The attitude of the issuer's management is one of the key determinants of such adjustment(s). The forward-looking statements within the textual contents of the prospectus are valid proxies of such attitude, which mostly fall in the *MD&A* sections of the prospectus (Hanley & Hoberg, 2010; Li, 2010). The forward-looking statements may involve certain topical contents, such as valuation, operation strategies, marketing, products/services, and so forth. On the other hand, the management's awareness of risks threatening the firm's performances is another key determinant of IPO valuations, according to prior finance studies (Campbell, Chen, Dhaliwal, Lu, & Steele, 2014; Clarkson & Thompson, 1990; Lewellen, 2006). Mostly, these mentions fall in the *Prospectus Summary* and *Risk Factors* sections – however, among all the mentions, only emphasized ones were investigated – i.e. mentions of risks with strong adverse adverbs – which highlight the management's awareness. This design decision is in accordance with a recent related study (Chan & Franklin, 2011). The goal of this research question is to analyze these topical contents in the forward-looking statements within the *MD&A* section, as well as the emphasized mentions of the risks in the *Risk Factors* section of the prospectus, and then assess their impacts on the IPO price valuation. The underlying premise is that more positive contents in the forward-looking statements indicate higher level of management's confidence; while more emphasized mentions of a certain type of risks indicate higher level of management's awareness towards the risks that the issuing firm encountered in the past and present or may encounter in the future.

1.3 Research Objectives

In this project, we propose an analytical framework for processing the textual contents within the IPO prospectuses in order to identify and extract the hidden semantic information related to indicators of IPO valuations, based on a particular approach of IE, namely *Ontology-based Information Extraction* (OBIE). With the help from the domain ontology, which is enhanced through automatic ontology learning techniques, not only the relevant concepts, but also the semantic relations between them and the attributes

defining them, are extracted from the original sources. Moreover, a reasoning and analytical engine is also be incorporated in the framework that enables further analyses toward the extracted knowledge for IPO pricing prediction purposes. A prototype system is then developed as the implementation of the proposed framework, which expands the contributions of this work from academia to practice. A thorough evaluation is also conducted on both the analytical framework and the prototype system in order to validate their functionalities, accuracies, and efficiencies.

There are two primary goals in this project. The first goal is to design and develop an ontology learning approach to enable (semi-) automatic construction of an underlying knowledge base for analyzing the informative contents contained in the IPO prospectuses. The second goal is to predict the IPO pricing volatility utilizing the information extracted via OBIE techniques, based on the domain ontology constructed in the first goal.

1.4 Scope and Contributions of the Study

The boundaries of this study include the process of knowledge extraction from the textual contents of IPO prospectuses and predicting price changes during (proposed) and after the IPO process. The two key components defining the scope of the study are automatic ontology enrichment and IPO pricing predictions – they are standalone yet logically inter-related. There are two key contributions made in this work. Firstly, the constructed domain ontology enables analysis toward the IPO prospectuses, as well as other similar financial reports. The techniques developed for constructing the ontology can be used in other domains (i.e. healthcare, social network analysis) as well. Second, we design and develop a decision support system via presenting IPO pricing trends to the investors, in order to assist the investment decision-making process – which ensures the practical relevance of this work. These two research contribution components are elaborated below.

The ontology enrichment approach proposed in this work prepares the knowledge foundation of the IPO pricing analytics. We design and implement an ontology learning technique for enriching a full-fledged domain ontology, as a specification of the domain

knowledge base, from the textual contents in IPO prospectuses. Ontologies are recognized as formalized repository of domain knowledge, which can also be used as extraction guides, reasoning basis, and representation schemes within an OBIE process (S. Sen, Tao, & Deokar, 2014). However, manually constructing and enriching ontologies are demanding and tedious activities (Zhou, 2007). In order to overcome such shortcomings, automated learning techniques, as a gateway to (semi-) automatically create or extend ontologies using machine-learning techniques, have been well studied. Information Extraction (IE) has been proven to be one of the most important techniques in terms of enriching ontologies. Among the various sources for IE, text mining turns out to be one of the most effective approaches (Wong et al., 2012). An important application of text mining in that regard is to extract and disambiguate relevant terms from domain-specific corpus, as well as to identify relations, for the purpose of enriching knowledge bases – and ultimately support the knowledge acquisition and subsequent activities in the knowledge management process. Even though several automated enriching approaches have been developed in recent studies (Dorji et al., 2010; Gaeta, Orciuoli, Paolozzi, & Salerno, 2011; Ittoo & Bouma, 2013a; Kang, Delir Haghighi, & Burstein, 2014), current approaches rely highly on the size and the quality of the text corpora (as the *training set*) annotated against formal domain knowledge. Moreover, sometimes the extracted terms are vague or confusing, or too narrow to be used as a reasoning/analytical basis. Thus, a research gap exists in the area of term extraction, namely noise handling and knowledge richness in the ontology enrichment process. Also, non-taxonomical relation plays a significant role in ontologies (Jiang, Tan, & Wang, 2007; Sánchez, Moreno, & Del Vasto-Terrientes, 2012). Yet, existing literature treat term extraction and (non-taxonomical) relation extraction as separate or sequential processes. Studies on term extractions solely focus on taxonomical relations among extracted terms (Ittoo & Bouma, 2013b; Kang et al., 2014), while studies on semantic relation extractions assume that term extractions are independent. However, in reality they work in an interlocking manner and should be recognized in a synergistic fashion. Term extractions provide foundations for relation extractions, while relation extractions testify and enhance term extractions. As such, leveraging term and relation extraction in a systematic manner is another research gap that needs attention.

The major contributions related to this component are four-fold. First, we propose a novel approach that addresses aforementioned research gaps by synthesizing IE phases such as term extraction, domain specific term selection/filtering, Word Sense Disambiguation (WSD), and ontology integration/enrichment. The approach is aimed at learning relevant terms for updating a formal domain knowledge structure and emphasizes two major issues in terms of ontology learning, namely quality and efficiency. Also, the proposed approach adopts a feature-based method that assists in topic extraction and integration with existing ontologies in the given domain. Second, we present an innovative application of the proposed approach in the finance domain, particularly in the context of a corpus consisting of Initial Public Offering (IPO) prospectuses. A research prototype of this application is reported to illustrate the feasibility and the efficiency of the proposed method in understanding the Initial Public Offering (IPO) phenomenon. The case study intends to extend a manually created seed concept list with explicit and relevant terms extracted from a domain-specific document corpus. A preliminary empirical validation by the domain experts is also conducted to illustrate the accuracy and advancements of the proposed approach. The result from the case study indicates the advantages and potentials of the proposed approach. Third, we systematically incorporate two facets (term and relation extraction) of ontology enrichment process together, and discuss in detail how they work in tandem. Last but not least, we propose a novel set of evaluation metrics to deal with the limited nature of traditional quality measures that simply compare syntactic representations with taxonomy concepts rather than the semantics within these concepts.

Predictive modeling is the second key research contribution component within the scope of this study. We design predictive models utilizing the information acquired from the OBIE process, based on the ontology enriched via the ontology enrichment approach. The predictive models designed in this project are based on the linkage between key determinants, such as the management's confidence toward the organization's future performances, as well as its awareness toward the risks faced by the organization, and the pre-IPO offering price adjustment and the post-IPO initial returns. We use a sample of 513 completed U.S. IPOs with machine-readable Form 424B4 filings (one of the variant of Form 424) during the most recent decade (2003 – 2013) as the main dataset in this

study. The prospectuses in our sample are acquired using an incremental web crawler (developed by the author) from the SEC *EDGAR* database. We define two metrics as unbiased proxies of the key determinants represented in the textual contents of IPO prospectuses. To the best of our knowledge, this is the first study proposing these metrics to quantify management's confidence toward the organization's future outlook and its awareness of risks. Based on the enriched ontology, we build an analytical processing pipeline, which we call *FOCAS-IE* (*Feature Oriented, Context-Aware, Systematic Information Extraction*) to identify and quantify aforementioned metrics from Form 424 filings. Compared to prior related studies, such as (Ferris et al., 2013; Loughran & McDonald, 2013), *FOCAS-IE* analytical pipeline is different in two aspects: a) Features: *FOCAS-IE* extracts knowledge regarding aforementioned metrics, and isolates the impacts across different features specifically important in the *Risk Factors* and the *MD&A* sections. These features depict important aspects of the firm's risks as well as operational performance in the future (i.e. product market, corporation strategies, marketing, corporation governance, and valuation), and b) Context-awareness: Unlike traditional content analysis techniques, we extract and leverage sentiment information at the sentence level. The rationale behind this design is that we only focus on sentences highlighting certain prospect or risks in respective sections. For instance, in the *MD&A* section, we focus on sentences that discuss features, with respect to the future outlook (forward-looking), and with the sentiment signal words modifying these features, termed as "meaningful" sentences. Moreover, the proposed method also differs from traditional sentiment analysis by introducing a third "uncertain" sentiment in addition to the "positive/negative" dichotomy, given that uncertainty/neutral tone is an inseparable part of the tone of the management's confidence of the future outlook. The importance of uncertainty in the prospectus contents is highlighted by Loughran & McDonald (2013). In the *Risk Factors* sections, we focus on the sentences discussing risks related to key features (i.e. growth, management skills, competitiveness, customers, lawsuits, and stock price) that are emphasized through the use of strong modal adjectives/adverbs. Since the *Risk Factors* section discusses different types of risks faced by the organization, it makes sense to focus on emphasized risks that indicate the management's greater concern about these risks as compared to others. To our best knowledge, this is the first study utilizing a

feature-oriented, context-aware framework, particularly in the IPO valuation field. The insightful information extracted from *Risk Factors* and *MD&A* sections is then quantified and normalized through the implemented *FOCAS-IE* analytical pipeline.

1.5 Structure of the Dissertation

The remainder of the dissertation is structured as follows. Chapter 2 presents a literature review by focusing on the IPO valuation problem, as well as the developments in content analysis of IPO prospectus for understanding IPO pricing volatility. Current mainstream ontology learning approaches are also surveyed in this chapter. A set of design requirements, in response to the research gaps identified in the literature review, is articulated in this chapter. Chapter 3 discusses the research methodology and research plan in this project. The design science research methodology is followed and related guidelines are addressed (Hevner, March, Park, & Ram, 2004). Chapter 4 presents the design and demonstration of the analytical framework (which is one of the main design artifacts of this project), by discussing the key elements and associated techniques/approaches within it. Chapter 5 demonstrates and evaluates the design artifacts proposed in this work. The proposed framework is demonstrated in the *FOCAS-IE* analytical pipeline, and evaluated in a case study. Furthermore, in order to ensure the practical relevance of the study, we design a set of experiments, in which the knowledge extracted from the textual contents in the IPO prospectuses, are used in predictive models, in order to predict the pricing trends prior and subsequent to the IPO date. Chapter 6 concludes the report by summarizing the contributions of the study. The challenges, implications, and limitations of the study are discussed in this chapter as well. Suggestions of the future steps are also presented in this chapter.

CHAPTER 2

LITERATURE REVIEW

In this section, we provide an in-depth literature review related to various topics in the scope of this study. First, recent studies on IPO valuations are surveyed, followed by highlighting limitations in these studies. In order to address the limitations of traditional IPO valuation studies, current research efforts are focused on investigating the informative contents within IPO prospectuses, ultimately aiming to understand IPO valuation in detail. These studies have been reviewed in Section 2.2, along with research gaps and areas of future work needed. In Section 2.3, *state-of-art* ontology learning techniques and approaches are reviewed critically. Section 2.4 summarizes the chapter by proposing the design requirements for the design artifacts in this study, based on the limitations of extant literature in this area and considerations of the scope of this study.

2.1 IPO Pricing Determinants: Theoretical Foundation

First, in order to demonstrate the relevance and significance of this study, we need to investigate current studies aimed at understanding IPO valuation strategies. A number of models and approaches focused on IPO pricing strategies have been proposed in the extant literature. Findings from these studies can be roughly categorized into two themes: *qualitative* and *quantitative*.

Traditionally, researchers have focused on the quantitative information (i.e. financial facts, numbers, and tables) disclosed prior to and during the IPO process. Regression and time-series analysis are two most popular methods used in the quantitative studies. Given the scope of this project, we reviewed IPO pricing models proposed based on the informative contents in the prospectus. For instance, Bhabra & Pettway (2003) have constructed a regression model to predict abnormal returns (less

desired pricing) with certain independent variables such as *relative offer size*, *number of risk factors discussed in the prospectus*, *degree of underpricing*, *underwriter reputation rank*, and a dummy variable *MGMT* (if the management sells shares at the IPO, $MGMT = 1$, otherwise, $MGMT = 0$). Further, Chemmanur (1993) have constructed a price estimation model based on information theory, in a scenario when a firm decides to sell stocks at both IPO and secondary markets. Also, a time-series based model is proposed by Ljungqvist, Nanda, & Singh (2006) to illustrate the correlations between investor sentiment and IPO pricing in a “hot” market.

Recently, researchers have also started to leverage qualitative information disclosed in the IPO process pertinent to IPO valuation. Qualitative methods may be used to estimate the trends of price changes in the IPO pricing (i.e. the stock price might increase/decrease), while the quantitative approaches may be used to reflect the actual changes in the pricing (i.e. the stock price might increase by 10%, etc.). As such, the quantitative methods provide more granularity (which in a sense is more useful if the investors are more concerned more about the actual change rather than the changing trends) but sacrifice the accuracy (it is much easier to estimate the trends than the actual changes).

Qualitative studies in this topic area identify several factors that impact the IPO pricing strategies (either *positively* or *negatively*) (Ritter & Welch, 2002). Studies that focused on the *underpricing* phenomena in an explanatory sense are proposed and are based on asymmetric information as well as symmetric information theories. Normally, IPO issuers/underwriters are more informed than the investors (they have resourceful insights into the *investor sentiment*, *growth opportunities* of the issuer, and *adverse selection considerations*, which are the major determinants of the IPO volume as well as pricing) and thus, they are more knowledgeable about the pricing processes (Lowry, 2003). On the other hand, according to the symmetric information theory, if both the underwriters and the investors are equally informed, the pricing (underpricing) could be attributed to the factors such as the *attitude of the underwriter* (optimistic/pessimistic), the *reputation of the underwriter*, the *perceived quality of the issuer* (often measured in the form of projected profit in the near future), *legal liabilities of the underwriters*, *industrial-dependent factors* (e.g., the technology industry after the Internet bubble), and

other rationales (i.e. providing compensations to induce potential investors, etc.) (Asquith, Jones, & Kieschnick, 1998; Carter et al., 1998; Jenkinson & Jones, 2009; Lowry & Schwert, 2004; Ritter & Welch, 2002; Roosenboom & Thomas, 2007). Other approaches used by studies in this category include: *lifecycle based theories*, and *evidence-based analysis*.

While these prior studies focused on IPO pricing strategies have provided valuable insights, some limitations may be noted as follows:

- With quantitative approaches, the rich qualitative information is ignored; thus, important indicators hidden in them, such as the management's confidence and prospect, are not considered in explaining IPO pricing strategies
- The initial attempts using qualitative methods resolved the issue to a certain extent. However, the quantification of the qualitative information is ad hoc and the analysis techniques are primarily manual, which requires domain expertise as well as intensive resources.

In order to handle above research gaps, researchers have shifted their focus to analyzing the qualitative/textual information within the IPO prospectus, which is a rich and very reliable information source within the IPO process (Hanley & Hoberg, 2010; Loughran & McDonald, 2013; Lowry & Schwert, 2004). A detailed review is provided below.

2.2 IPO Valuation and Informative Contents of Prospectuses

As discussed above, existing studies aimed at understanding IPO valuation and pricing strategy have reported results from (manual) investigations of quantitative, financial facts (majorly numeric) disclosed in the prospectus, particularly in the *Use of Proceeds* section (e.g., Ritter & Welch (2002), Bhabra & Pettway (2003), Jenkinson & Jones (2009), Löffler et al. (2005), Lowry & Schwert (2004)). In the past decade, computer-mediated content analysis has become a key direction for this research. For example, Arnold et al. (2010) examine the informative contents within the *Risk Factor* section of the prospectus, and make an argument that the ambiguity level of the disclosed information has a correlation with the pre- as well as post-IPO price volatilities. Their

findings suggest that more ambiguous information in the prospectus leads to greater price adjustment. Hanley and Hoberg (2010) conduct a content analysis on four most important sections of the prospectus, and classify contents within into *standard* and *informative* content. Their findings indicate that IPOs with more informative content (i.e. domain- and business-specific information) tend to have lower pre-IPO price revisions and post-IPO initial returns. The sentiment-oriented contents in the prospectus have also been analyzed, and they argue that positive tone related to valuation and due diligence (i.e. *accounting*, *corporate strategies*, and *products/revenues*) links to lower price revision and initial returns (i.e., market's price adjustment). In a follow-up study, Hanley and Hoberg (2012) further illustrate that the risks to future legal issues correlate to greater levels of underpricing of IPOs. These studies provide the foundation for our study by proving that content analysis is useful in understanding the IPO pricing phenomenon.

Few recent studies have used sentiments and tones expressed in the IPO prospectus textual contents to analyze IPO valuation. Loughran and McDonald (2013) analyze the tone of the prospectus (mainly *Form S-1*, which is claimed to be as important as *Form 424 filings*) and the IPO valuation by considering the occurrences of sentiment words from six sentiment word lists (the L&M word lists focused exclusively on the finance/IPO domain) developed by them in an earlier study (Loughran & McDonald, 2011) as a proxy for the overall tone of the prospectus. The major contribution of this study is that IPO prospectuses with more uncertain words (*negative* or *weak modal*) have higher pre-IPO price revisions and post-IPO initial returns. In a similar study by Ferris et al. (2013), the authors proxy the management's confidence from the perspective of conservatism, which is calculated as the ratio between the negative words and total words in the prospectus. The study also utilizes the L&M word list for analyzing sentiments; and the findings in their paper are consistent with the study by Loughran and McDonald (2013). Besides the IPO prospectuses, other knowledge resources have also been reported in examining IPO valuations or management's confidence. For instance, in studies such as (F. Li, 2010; M.-C. Lin et al., 2011), annual and quarterly corporate financial reports are used (i.e. 10-K, 10-Q) for content analysis to understand the management's prospects and stock price movement, respectively. Other studies utilize the financial news and social media contents as the knowledge resources (Hagenau, Liebmann, & Neumann,

2013; Schumaker & Chen, 2009). When comparing to IPO prospectuses, these sources lack comprehensiveness, authority, and compliance with respect to the IPO process.

While the aforementioned studies report substantial progress, following research gaps may be noted: (a) Arguably, without the consideration of the context, mere counts of the sentiment-oriented words are not precise proxies of the management's outlook. For instance, a simple negation in the context might change the direction of the sentiments. Similarly, phrases expressed in a certain structure may imply a different meaning. Thus, a context-aware sentiment analysis is deemed necessary in analyzing prospectus contents. Even from a generic sentiment analysis standpoint, this advancement can be further adapted to various domains. It is also noteworthy that most of the surveyed studies posit on an explanatory standpoint, while it is clear that predictive analysis might yield higher relevance to the domain. (b) As indicated in Hanley and Hoberg (2010), the topical contents disclosed in different sections are different – thus, their impacts on the IPO valuation might vary. On the other hand, even the same sentiment on different features might have different effects on the IPO pricing (e.g., a negative sentiment on marketing and a negative sentiment operation strategies could have opposite effects on price revisions). Thus, a section-by-section, feature-oriented analytical approach would be worthwhile. (c) Existing studies lack an underlying knowledge structure that can serve as a basis for reasoning and analytics (e.g., considering features and contexts). Thus, a well-constructed (in terms of coverage, accuracy, and domain specificity) ontological structure is needed that can be leveraged for analytical tasks on IPO prospectus content.

2.3 Ontology Enrichment from Textual Corpus

Several previous studies have proposed term extraction methods in various domains (Dorji et al., 2010; Gaeta et al., 2011; Ittoo & Bouma, 2013b; X. Jiang & Tan, 2005; Y.-B. Kang et al., 2014; Q. Li & Wu, 2006; Nassif et al., 2009; C. Zhang, Niu, Jiang, & Fu, 2012). Only two studies (X. Jiang & Tan, 2006; Y.-B. Kang et al., 2014) have used domain specific ranking mechanisms for selecting domain-specific terms. Moreover, none of the method employed a domain specific cleansing step to further filter the extracted terms, which is essential for data quality. Even though some of these

methods achieve better accuracy, it is partially because some of them utilize a strictly tested domain ontology (Kang et al., 2014) and/or supervised methods with a training data set (Nassif et al., 2009). A related topic area on this context is that of *Word Sense Disambiguation* (WSD). A number of studies have been done in developing/improving WSD methods – a detailed review can be found in Wong et al. (2012). However, a major limitation of these studies is the need for voluminous text corpus to extract reasonably accurate relations and the general focus on higher level of granularity (e.g, sentence-level) as opposed to finer granularity (e.g., clauses). There is a rising need for a better WSD-based approach that can be applied in domains where no explicit, formal knowledge structure exists (such as the IPO domain in this study); or cost-sensitive domains where obtaining a (large) training data set is not feasible. Such an approach may serve as a pilot step in an iterative term extraction process or to conduct other ontology-based analyses such as *information extraction*, *document categorization*, or *ontology-based reasoning*. Prior related studies have applied traditional Information Retrieval (IR) based metrics, such as *precision/recall/F-score* (Kang et al., 2014; Punuru & Chen, 2011; Shen, Liu, & Huang, 2012), which are incapable of identifying the semantic capabilities among terms. As such, there is a distinct need for a set of metrics for evaluating the quality of term extraction. Meijer, Frasincar, & Hogenboom (2014) is the only study we found that proposes and incorporates metrics related to terms. However, since this study focuses on term extractions, only taxonomical relations are considered in proposed metrics. Thus, metrics that can also consider non-taxonomical relations (both pre-defined and extracted) are needed.

2.4 Design Requirements

Existing studies reviewed earlier have helped articulate the research gaps and the need for a text analytics-based decision support framework to assist decision makers during the IPO process. Based on these, a set of requirements has been articulated aimed at the design and development of a text analytics system.

The proposed framework and associated analytics system should:

- **Provide a formalized framework containing relevant domain knowledge:** IPO documents (i.e. prospectuses) embed domain knowledge of distinct aspects, such as *organizational descriptions*, *managerial information*, *competitiveness*, *business policies*, and *expectations of the future* (Hanley & Hoberg, 2010). A normalized conceptualization should be created as a vault for such information in order to provide references/guidelines for the analyzing activities. This conceptualization should also be able to evolve by adding newly-discovered knowledge into it (semi-) automatically in an iterative fashion.
- **Identify occurrences of relevant concepts/entities regarding the IPO process:** Prospectus documents might contain several different expressions (i.e., “customers” and “buyers” might refer to the same entity) or different parts of speech (i.e., “issue” and “issuance”) of the same term. These different mentions of the same entity should be identified and then associated with the belonging concept with its instances correctly. Further, the extent of the mentions (i.e., the span of the representation of a particular entity in the text) should also be determined to reduce information redundancy.
- **Discover the relations between extracted concepts:** Different types of relations may exist between entities in the textual content of the IPO documents. For instance, a term such as “granting award” might be a particular type of another term such as “employee benefit plan”. Sometimes such relations are not explicitly defined in the textual content. Thus, the proposed framework should be able to extract such relations.
- **Extract hidden semantic information in the textual contents of the prospectus documents and use them for analytical/reasoning purposes:** Other than relations mentioned earlier, other types of linkage (i.e., *co-dependency*, *causal links*, etc.) may also exist between entities in the form of patterns. The proposed framework should be able to detect such patterns and recognize the linkages within them through analysis of the implicit semantic information hidden in the texts. For instance, “finance incapability” is often the reason of “failure to raise capitals”, which indicates a potential causal link between the two terms.

- **Enable semi-automated analysis by decreasing human interventions:** As stated earlier, current studies rely highly on manual efforts in the analytic phases. For instance, the analytical rules are encoded solely by domain experts. In the proposed framework, the extracting process should be supported in a semi-automated manner.
- **Support contextual analysis of IPO documents:** An IPO prospectus may be roughly divided into several sections (i.e. *Prospectus Summary*, *Risk Factors*, *Management Discussion & Analysis*, etc.). The appearances of the same term in different contexts (i.e. *section*, *subsection*, *paragraph*, *sentence*, etc.) could imply different semantic meanings. As such, the proposed framework should support both Delta analysis and contextual analysis.
- **Present the discovered knowledge in a normalized form and enable user-defined querying:** Besides abovementioned requirements regarding the knowledge discovery process itself, how to present the results from such process to the users would be another important issue. The proposed framework should support presenting output in the form of the normal conceptualization, allowing users to cast queries for desired question-answering purposes, and should support automatically adding the extracted knowledge to the domain knowledge base through mechanisms such as ontology instantiation, rule integration, and ontology learning.
- **Provide insights regarding IPO valuation from a predictive standpoint:** To ensure the practical relevance of the study, the knowledge acquired from previous steps/requirements need to be used in analyzing the IPO pricing/valuation phenomena. A method need to be developed in order to quantify such information representing certain implicit knowledge (i.e., management's awareness of risks, management's confidence of future performances), and then apply them in proper predictive models, for IPO pricing prediction purposes.

Based on the design requirements mentioned above, we propose the analytical framework in Chapter 4. In the next chapter, we first review the research methodology that has been adopted in this research study.

CHAPTER 3

RESEACH METHODOLOGY

This chapter discusses the research methodology employed in this dissertation. In this work, this study employed the design science research methodology, which is elaborated in Section 3.1.

3.1 Design Science Research Methodology

As discussed earlier, there is a need for new and innovative analytical tools and techniques to analyze the IPO pricing phenomena. Since the design problem is well outlined in Chapter 1, and the requirements of a solution are articulated in Chapter 2, in this Section, we intend to discuss the design science research methodology adopted in this work, along with the design artifacts, following the guidelines proposed by Hevner et al. (2004). Each of these guidelines is discussed below in the context of this study.

3.1.1 Design as an Artifact

As discussed in the paper by Hevner et al. (2004), four major types of design artifacts are expected as research outputs from a design science research project, namely: construct, model, method, and instantiation. The key design artifacts developed in this study are as follows. First, building on existing constructs in the context of financial domain and information extraction techniques, this research proposes artifacts in the form of a “method”. The proposed analytical framework (discussed in detail in Chapter 4) and the ontology reasoning and learning method presents a novel method for analyzing textual contents of IPO prospectuses aimed at pricing prediction goals. Second, the enriched IPO Ontology, as an intermediate outcome of the study, contains domain-specific terminologies and semantic relationships among these terms and can be viewed

as a “model” from a design artifact standpoint. Additionally, predictive models built for pricing prediction using textual features/sentiments extracted from IPO prospectuses are another set of “models” produced from this research study. Last, but not least, “instantiation” of the analytical framework in the form of a prototype system and *FOCAS-IE* analytical pipeline are also design artifacts from this study.

3.1.2 Problem Relevance

The outcome of design-science research in the IS domain should assure importance and relevance of solving business problems (Hevner et al., 2004). The relevance of this study is two-fold. On one hand, due to the topic and scope of this dissertation project, it is relevant to the finance and accounting domain. As stated earlier, understanding IPO pricing has significant values in the finance domain, from both the academic and practical perspectives. Moreover, the knowledge base constructed in the process (*IPO Ontology*), can be used as knowledge repository and/or reasoning basis in the domain of accounting. On the other hand, this study will contribute to the IS domain knowledge base by proposing approaches, methods, techniques, and metrics to IS sub-domains such as text analytics, predictive modeling, as well as ontology learning. This could possibly lead to further studies in a variety of additional problem solving processes, such as healthcare, terrorism detection, and so forth.

3.1.3 Design Evaluation

To ensure the rigor and relevance of the design artifact, well-designed evaluation steps need to be carried out to validate its functionality, accuracy, and efficiency (Hevner et al., 2004). The design artifacts proposed in this work are validated thoroughly, in a two-phase evaluation process. Firstly, each of the design artifacts is evaluated independently. For instance, a “*ground-truth*” based method is adopted in the evaluation of the ontology learning approach proposed in this work, while the evaluation using empirical data is applied for the predictive modeling approach. Furthermore, at the framework level, the predictive modeling is considered as part of the evaluation for the prior modules, while the developed research prototypes also demonstrated the functionalities and efficacies of the proposed approaches. A combination of different

evaluation methods is applied in this work. The underlying theoretical framework for evaluating design artifacts in this work can be found in Chapter 5.

3.1.4 Research Contributions

The outcome(s) from design science research must be clear and verifiable, with respect to design artifacts, foundations, and methodologies (Hevner et al., 2004). As stated in Section 2.4, the contributions of this work are clear and valid. For instance, the analytical framework provides an effective approach for predict IPO Pricing using domain-specific knowledge, while this knowledge could also be used for other purposes (i.e. fraud detection). Further, the analytical framework itself can be extended to other domains, such as healthcare, e-learning, and so forth. The design of the framework is based on existing IE literature, while the analytical basis (*IPO Ontology*) is constructed based on findings from existing finance studies. A top-down design methodology is followed in this work. All of the design artifacts (by the means of the prototypes and results), design foundations, and design methodologies are verifiable.

3.1.5 Research Rigor

Design science research demands rigorous construction and evaluation methods in terms of the design artifacts (Hevner et al., 2004). A variety of mechanisms are followed to assure the rigor of this study. First, the effectiveness of using knowledge base is tested in this study. For instance, in the ontology learning module, the similarity between the automatically constructed ontology and the manually constructed ontology (for the purpose of serving as the domain knowledge base) are compared. Complexity, performance, and usability metrics are developed and used in evaluating design artifacts. More details of these validations can be found in Chapter 5.

3.1.6 Design as a Search Process

The goal of design science research is to search for “the best, or optimal design, which makes the design science projects inherently iterative” (Hevner et al., 2004). An iterative design strategy is closely adopted in this work. The final design artifacts of this study have evolved largely by making iterative comparisons to the earlier designs. These

evolutions have improved the ontology learning capabilities as well as capabilities of predicting IPO pricing. Furthermore, based on the proposed approach in this work, additional research opportunities have revealed themselves. Through future round(s) of iterations, the capabilities, accuracy, as well as usability of the design artifacts will be enhanced to solve more design problems.

3.1.7 Communication of Research

The last guideline requires design science research to be presented in a way that is understandable for both technology- and management-oriented audiences (Hevner et al., 2004). Due to the interdisciplinary nature of this study, communication of this study is very crucial. To ensure this point, I have worked closely with experts from the finance domain, conducted intensive literature review in the finance domain, and am prepared to repackage the results from this work and publish it in the finance domain; so that the domain/practitioner relevance can be ensured.

CHAPTER 4

THEORY AND ARTIFACT DESIGN

In this chapter, I first demonstrate the architecture of the proposed framework in Section 4.1. Then, the three major modules are discussed in subsequent sections. Section 4.2 discusses the Information Extraction module in the framework; Section 4.3 demonstrates the Reasoning and Learning Module; while Section 4.4 presents the Analytics Module. Instantiation details of different modules are discussed in respective sections.

4.1 An Overview of the Analytical Framework

Drawn on the design requirements highlighted in Section 2.4, the design of the analytical framework is depicted in Figure 4.1 below.

The relevant information within the prospectus documents are annotated by a text-processing application built on a NLP platform namely “General Architecture for Text Engineering” (GATE). GATE is a comprehensive architecture supporting NLP engineering and providing IE implementations such as *tokenizers*, *sentence splitters*, *part-of-speech taggers*, *gazetteers*, *pattern-matching grammars*, *stemmers*, and *co-reference resolution* (Cunningham, Maynard, Bontcheva, & Tablan, 2002). Some of these implementations have been modified to support the information extraction module in our prototype system. A grammar rule language named JAPE (Java Annotations Patterns Engine) is used in the preprocessing, annotation, and the ontology population activities. Essentially, JAPE rules are finite-state transducers. Each transducer contains a different type of rules, while each rule has a Left-hand Side (LHS) and one (or multiple) Right-hand Side (RHS). The LHS is used for pattern-matching while the RHS could be used for annotation and other aforementioned purposes (Cunningham, Maynard, & Tablan, 2000).

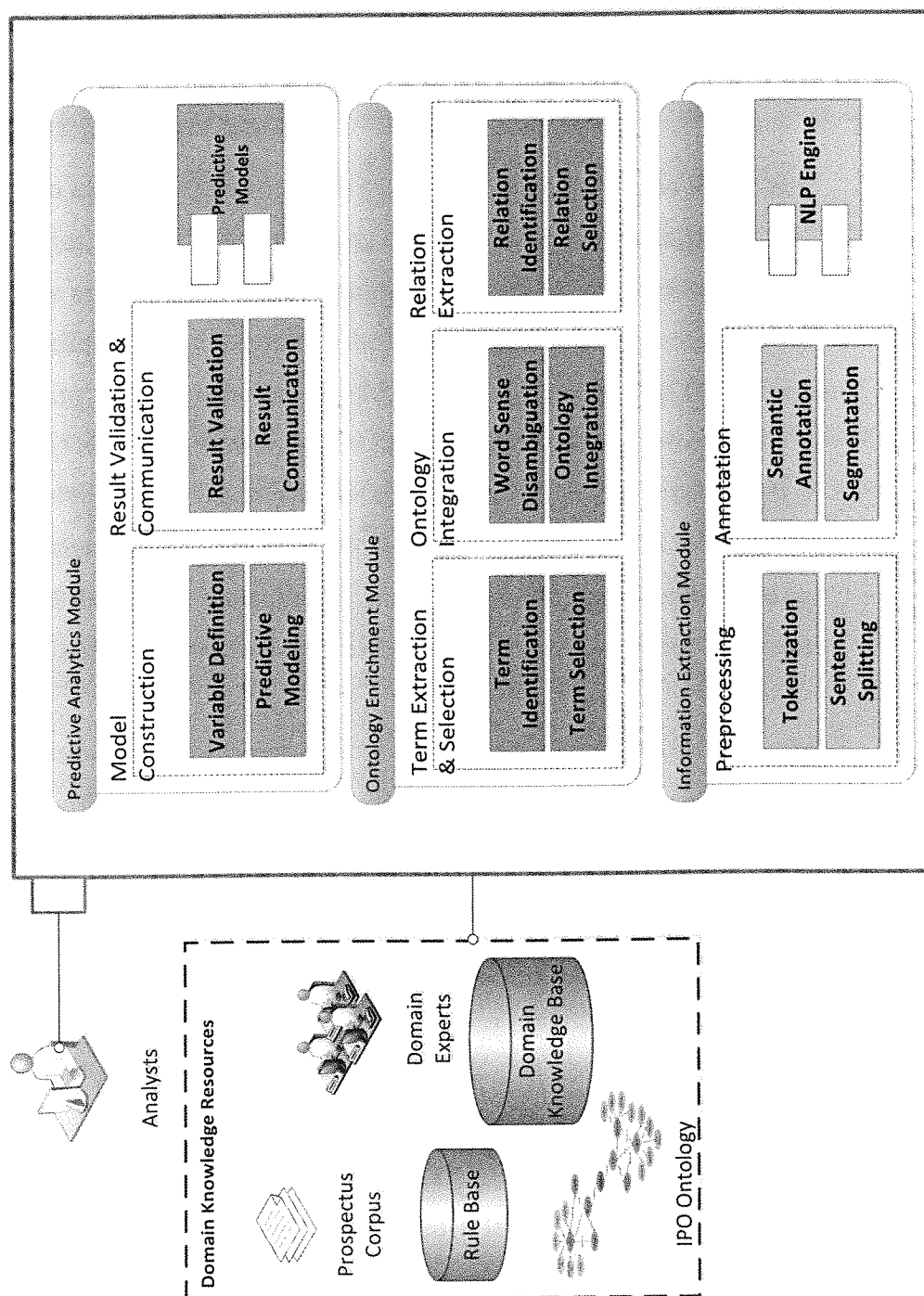


Figure 4.1. The Architecture of the Proposed Analytical Framework

As shown in Figure 4.1, the analytical framework contains three modules, namely information extraction module, ontology enrichment module, and predictive modeling module, respectively. Each of the modules and related components are described below.

4.1.1. Information Extraction Module

The information extraction module, with an information extractor as the kernel, contains several activities, such as: pre-processing the prospectus, annotating the prospectus based on the pre-defined domain ontology, and populating the ontology with extracted entities and relations. With respect to the design requirements mentioned earlier, the pre-processing of the prospectus should contain segmentation of the document (for contextual analysis) and change tracking (for *Delta analysis*, etc.). Further, the extractor provides capabilities such as prospectus entity recognition, entity relation recognition, co-reference resolution, and so forth, as shown in Figure 4.1. One thing worth noting is some of the sub-module(s) in the Information Extraction module are re-used in other two modules. Details regarding these sub-modules can be found in Section 4.2.

Text pre-processing is one of the most important phases in most IE systems, and it provides annotations and other information for subsequent use. Particularly in our project, document segmentation and change tracking are two of the most significant activities in the pre-processing stage. In this case, different JAPE rules are coded and stacked as a pipeline in order to facilitate these activities.

4.1.2. Ontology Enrichment Module

The Ontology Enrichment module synthesizes IE phases such as term extraction, domain specific term selection/filtering, Word Sense Disambiguation (WSD), and ontology integration/enrichment. The approach is aimed at learning relevant terms for updating a formal domain knowledge structure and emphasizes two major issues in terms of ontology learning, namely *quality* and *efficiency*. Also, the proposed approach adopts a feature-based method that assists in topic extraction and integration with existing ontologies in the given domain. Second, we present an innovative application of the

proposed approach in the finance domain, particularly in the context of a corpus consisting of Initial Public Offering (IPO) prospectuses. Furthermore, this module also enables identification and extract semantic (non-taxonomical) relations from textual contents. If the Information Extraction and Predictive Modeling modules utilize the ontological information from the IPO Ontology, this module fulfills the other half of the iteration – which updates the ontology with formalized, domain-specific information in real time. The details related to this module and its sub-modules are provided in Section 4.3.

4.1.3 Predictive Analytics Module

The third module in the analytical framework is the predictive modeling module, which utilizes intermediate results from the other two modules, via predictive modeling techniques, to predict IPO pricing trends. Prediction is certainly the main functionality in this module; however, according to Section 3.2, results validation and communication are also part of this module. Users are able to verify the results and associated modeling techniques via accuracy and performance metrics, while interpretation assistance and data visualization tools help the communication of the prediction results. Details related to this module are provided in Section 4.4.

4.1.4 IPO Ontology

The analytical framework, as an OBIE application, relies largely on an ontology as the underlying knowledge structure. The IPO Ontology is involved in this analytical framework in an iterative fashion: the IPO Ontology provides the extraction guidelines and reasoning basis for aforementioned modules, while these modules, in turn, enrich the IPO Ontology on the fly. Domain experts construct the initial version of the ontology, then with ontology enrichment techniques (discussed in Section 4.3), extracted and filtered information from the IPO prospectus corpus are added to IPO Ontology in a semi-automatic fashion. The structure of IPO Ontology is shown below (Figure 4.2). For better readability, instances (used for annotations) and detailed semantic relations (used for reasoning) are omitted in Figure 4.2.

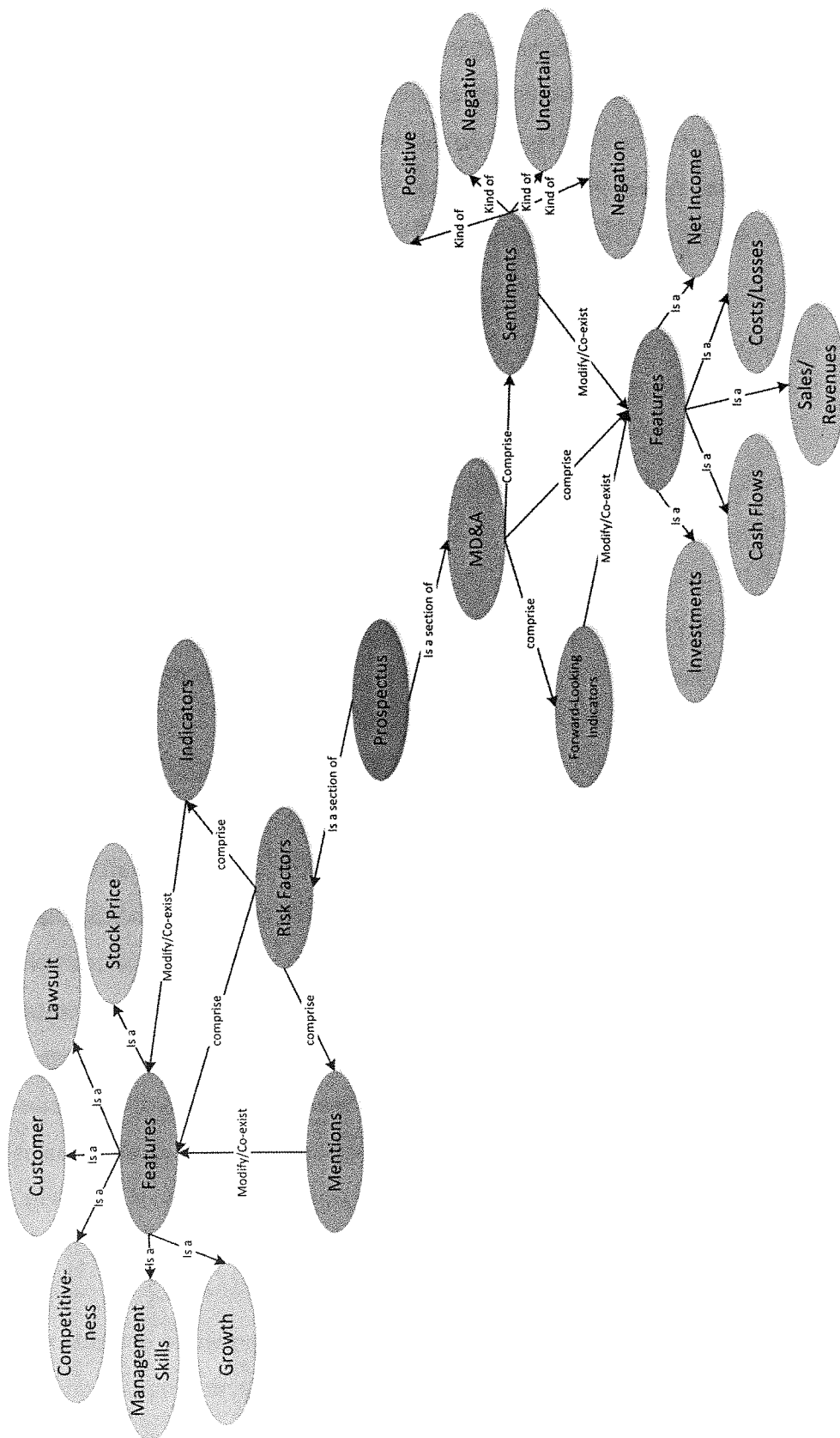


Figure 4.2 Overview of IPO Ontology

As discussed above, IPO Ontology serves as the kernel of the proposed analytical framework. The Information Extraction module relies on the terms and terminological relations in IPO Ontology; the Ontology Enrichment module populates and updates IPO Ontology with discovered knowledge elements; the Predictive Modeling module recognizes IPO Ontology as the knowledge base for reasoning and subsequent prediction purposes. Details regarding the application of IPO Ontology in relation to each module can be found in relevant sections below.

4.2 Information Extraction Module

The Information Extraction module is a key module in the proposed framework. Some of its functionalities, such as text pre-processing, lemmatization, and so forth, are reused in other modules. Two steps exist in the Information Extraction module, namely text preprocessing and annotation.

Typical preprocessing steps/tasks in an IE system/module include: tokenization, sentence splitting, POS tagging, stemming, parsing (removing stop words and “useless” elements such as XML/HTML tags, and other document preparations), and so forth. These tasks are usually conducted with the help of an NLP platform (i.e., GATE). A detailed discussion toward these steps can be found in a research article by Appelt (1999). Hereby we are discussing the design of the Information Extraction module in our framework in following paragraphs.

The segmentation of the documents enables the later contextual analysis. In this phase, contextual information and the extracted knowledge is added to the IPO ontology. The JAPE rules for such function first identifies the section titles and then the section contents – then the relative position of the extracted entities is determined. The prospectus contents in the prototype are identified at the section level (i.e. *prospectus summary*, *risk factors*, etc.). The implementation of such functions in GATE is shown in Figure 4.3.

The semantic annotation of the prospectus documents, following the phases of entity recognition, relation recognition, and the attributes identification, is implemented in GATE. The ontology-based gazetteers (provided within GATE) in combination with a

plug-in to GATE, namely APOLDA (*Automated Processing of Ontologies with Lexical Denotations for Annotation*), have been applied to provide semantic annotation of documents following the paradigm of lexicon-based annotation (Wartena, Brussee, Gazendam, & Huijsen, 2007). The APOLDA plugin is more suitable dealing with a large amount of concepts (classes) with less textual representations, while the ontology-based gazetteers are more useful when the ontology has fewer classes. The results of the semantic annotation of the prospectus document are shown in Figure 4.4.

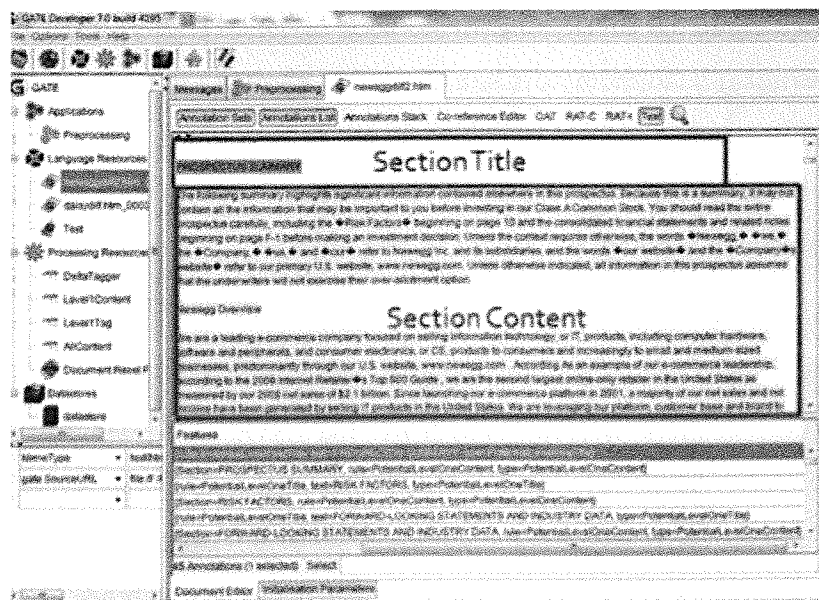


Figure 4.3. Segmentation of IPO Prospectus

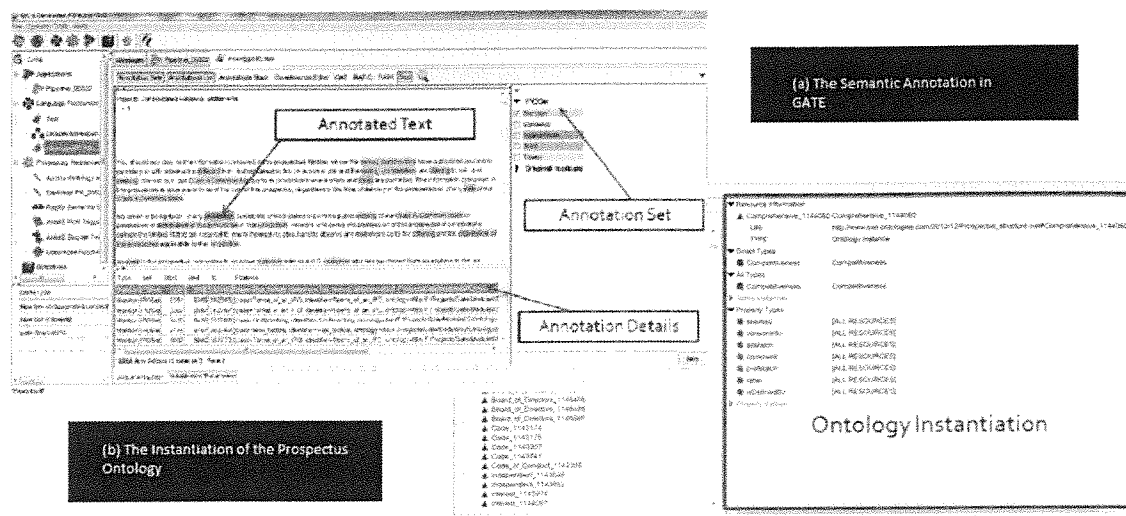


Figure 4.4 Semantic Annotations and Ontology Population/Instantiation

The semantic annotations of the discovered entities consist of several characteristics (termed as features in Figure 4.4) such as: the name of the belonging ontological class, the URI of the ontological class, the span of the annotation (in the form of *start* and *end* nodes), annotated texts, the locations of the texts (i.e. in which sections, paragraphs, etc.), and the change type of the text if available (i.e. *added*, *removed*). Another set of JAPE rules is coded to populate the Prospectus Ontology by adding the annotations (in the form of ontology instances) and aforementioned features (in the form of *datatype* properties). The output of the instantiation process is depicted in Figure 4.4 below. The left part of Figure 4.4 includes the instances identified from the semantic annotations in the prospectus. The right part of Figure 4.4 includes documents ontological classes (pre-defined, same as those in Figure 4.2) and ontology instances.

4.3 Ontology Enrichment Module

The proposed approach is to enrich domain ontologies from domain-specific textual resources. As stated above, ontologies can be used as a formal conceptualization for annotating, querying, reasoning, and other analytical purposes. The accuracy and efficiency of ontology-based analyses relies on the quality of the ontology, namely coverage and clarity. Coverage refers to the completeness of the ontology – i.e. the amount of terms/relations formulated in the ontology, while clarity refers to the explicitness and lucency of the terms in the domain ontology. The proposed approach is aimed at improving both coverage and clarity of an ontology: for increasing the coverage of the ontology, a mechanism is designed to extract and filter related domain-specific terms from a document corpus in a certain domain; for improving the clarity of the ontology, a Word Sense Disambiguation (WSD) method is proposed to reduce the conceptual confusion of the selected terms. The approach includes a mechanism for aligning the newly discovered terms with existing ontologies. Finally, a novel method is incorporated with the proposed approach to derive non-taxonomical relations between

extracted terms from same text corpus. In following sub-sections, we discuss these aspects in detail.

4.3.1 Domain Specific Term Extraction and Selection

Researchers have indicated in the literature that noun phrases in texts are roughly term candidates in most cases (Gaeta et al., 2011; Ittoo & Bouma, 2013b; Y.-B. Kang et al., 2014). Generally, the term extraction and selection process follows three schools of approaches: 1) corpora based approaches; 2) heuristic approaches; and 3) hybrid approaches. Corpora based approaches utilize the *Part-Of-Speech* (POS) tags and syntactic patterns provided by Natural Language Processing (NLP) tools. An example of the linguistic patterns can be found in the article by Li & Wu (2006): noun phrases match the syntactic pattern of $(JJ)^*(NN)^+$ are selected as candidates from parsed documents (where **JJ** refers to an adjective, **NN** denotes to a noun, * denotes zero or more occurrences (optional), while + denotes one or more occurrences (required)). The drawback of linguistic based approaches is that commonly their results rely largely on the amount of cross-sectional documents in a corpus. Thus, applying these approaches in a less mature domain will possibly result in poor outcomes. On the other hand, heuristic approaches rely on the frequencies/statistical measures of noun phrases extracted from the document collection. One of the most important measures is *term-frequency-inversed-document-frequency* ($tf*idf$), and its variants (one of them is Aizawa (2003)). Despite the importance and usefulness of aforementioned measures, they are not directly applicable to the current research project; the reason is similar to the discussions in Jiang & Tan (2005), terms with low $tf*idf$ scores perform better in domain specific cases. A hybrid based approach is a combination of the former two types of approaches (Kang et al., 2014).

In this paper, we develop our approach along the lines of a hybrid approach. The approach matches pre-defined linguistic patterns for term candidate selection and utilizes a domain specific heuristic measure for term filtering. Firstly, we have expanded the aforementioned linguistic pattern for our domain specific term extraction purpose. Note that English language stop words are removed from the term candidates, yet determiners (i.e. *a, an, the*) are kept for pattern matching purpose. In the current phase of this project,

we only capture the noun phrases from the document corpus. The *noun phrase patterns* (NPPs) captured in regular expressions (along with examples) are reported in following table (Table 4.1).

Table 4.1. Noun Phrase Patterns

NPP	Example
(DT)* (JJ)* (NN)*	legal proceedings, profits
(NN)* (IN)* (NN)*	strategy of competitors

In Table 4.1, DT denotes determiners, while IN denotes prepositions. One point worth noting is that the first NPP has two forms: single-word terms (with only one **NN**) and multiple-word terms.

With the term candidates extracted, the proposed approach calculates the filtering measures of term candidates in the specific domain through information along two different lines: heuristic information and domain-specific information. For the heuristic information, we propose a ranking measure as shown in Equation (1).

$$rank(t, d) = \sum_{i=1}^{|t|} \frac{freq(n_i, d)}{\max[freq(t, d)]} \times \log\left(\frac{df(n_i)}{\max[df(t)]} + 1\right) \quad (1)$$

In Equation (1), t is an extracted term candidate, $|t|$ is the number of nouns in t , n_i is the i^{th} noun in t . $freq(n_i, d)$ is the occurrence of n_i in document d . Given TC is the set of all term candidates ($t \in TC$), $\max[freq(t, d)]$ is the highest occurrence in d ($\forall t \in TC$). $df(n_i)$ is the occurrence of n_i in a (domain specific) glossary. If none domain specific glossary exists, then WordNet (Princeton University, 2012) is used as a domain-independent glossary. $\max[df(t)]$ is the maximum of the occurrence of any t in TC that appears in the glossary. The first part of Equation (1) represents the frequency of the term ($FREQ$), while the second part of it represents the domain relatedness of the term (DR). With the term candidates sorted based on the ranking measure, users can define the amount of terms needed. For instance, if the user decides to select 100 terms, then the top 100 terms from the sorted list are selected. Alternatively, the user can define a threshold on the ranking measure. For example, if the threshold is 0.6, then any term with $rank(t, d) > 0.6$ is selected.

Moreover, a deep cleansing step is incorporated in the term selection phase in order to enforce the domain relatedness of selected terms. All terms that meet the aforementioned ranking measure are further filtered through such rules. These rules are encoded based on the analytical purposes based on the terms. For instance, if the terms regarding the geographic locations are not relevant in further analysis, a rule will be encoded and enforced as: `DROP (NP(Token.category = "NN" && Token.kind = "LOC"))`.

4.3.2 Word Sense Disambiguation

Word Sense Disambiguation (WSD) is a computational process to identify the explicit meaning of words in a certain context (Navigli, 2009). WSD is an *AI-complete* problem, which means it is among the most difficult problems in the artificial intelligence domain. WSD can enhance the learned ontologies by reducing the terminological confusion within them (Wimmer & Zhou, 2013). Generally, there are two approaches for WSD: a) learning-based approach; and b) knowledge-based approach. Learning-based approach can be further categorized into *supervised learning based WSD* and *unsupervised learning based WSD*. A key difference between the two is *supervised learning based WSD* relies on tagged documents as a training set for future learning, while unsupervised learning does not require a training set; thus, even though it yields better results, it requires pre-tagged training set – which are usually not available in a less-well-defined domain or on a large sample size. Considering the nature of this project, particularly the size of the documents in the corpus and the domain expertise required to constructing a training set, adopting supervised learning based approaches is not feasible.

Knowledge-based approach can be further grouped into *dictionary-based* approaches, *corpus-based* approaches and *social media based* approaches. *Dictionary-based* approaches rely on external lexical resources, such as machine-readable dictionaries, thesauri, and ontologies (i.e. WordNet), whereas *corpus-based* approaches do not use any of them, rather than the empirical information of the textual contents. Instead of using dictionaries, *social media based* approaches utilize web content (i.e. Wikipedia) as the knowledge base, however, the data quality of such online sources is not

guaranteed. In this project, we believe the *dictionary-based* approach is better than the other two.

Before we present the WSD algorithm, we need to discuss the structure of WordNet (Princeton-University, 2012). We utilize the WordNet taxonomical relations for disambiguating word senses: basically, children classes of current term as hyponyms, parent classes as hypernyms, and sibling classes as synonyms. In order to simplify the computation complexity, we limit the scope to direct parent/children classes only.

We present a *feature-based* approach in this paper. Two types of features are adopted in this work, namely *local features* and *syntactic features* (Navigli, 2009). Local features represent a small amount of words around the target word, which their properties such as POS tags, word forms, positions; whereas syntactic features represent syntactic information related to the words surrounding the target word. The difference between local features and the syntactic features is that local features are *n-gram* bag-of-words (*n* words surrounding the target word), while syntactic features are features of the words within the same linguistic unit (phrases, sentences, paragraphs, etc.). Words for syntactic features might be outside the *n-gram* bag-of-words. Only words with POS tags of **NN** (nouns), **VB** (verbs), and **JJ** (adjectives) are considered as target words. The feature-based WSD (*F-WSD*) algorithm is presented in Figure 4.5.

Algorithm 1: Feature based WSD (F-WSD)

INPUT: BagOfWords = $\{w^a, W^d\}$ -- w^a : target word, $W^d = \{w_i^d\}$: set of surrounding words
INPUT: WordNet -- contains the WordNet ontology and related functions
OUTPUT: (w^a, sense_i) -- assign an explicit sense_i to w^a

```

1: //initialize
2: //Define Variables
3: domain_Sense_List{ $\text{sense}_i$ }; //WordNet senses of  $w^a$  tagged in the specific domain
4: hyper_Lista, hypo_Lista, syn_Lista; //hypernyms, hyponyms, and synonyms of  $w^a$ 
5: hyper_Listd, hypo_Listd, syn_Listd; //hypernyms, hyponyms, and synonyms of every  $w^d$ 
6: GET domain_Sense_List FROM WordNet;
7: //BEGIN
8: FOREACH ( $\text{sense}_i$ : domain_Sense_List){
9:   GET hyper_Lista, hypo_Lista, syn_Lista FROM WordNet;
10:  FOREACH ( $w_i^d$ :  $W^d$ ){
11:    IF(hyper_Lista, hypo_Lista, syn_Lista CONTAINS  $w_i^d$ ){ //if surrounding words
12:      RETURN this.sensei; //appear in hypernyms, hyponyms, or synonyms of this
13:      BREAK;} //sense, select this one
14:    ELSE IF((hyper_Lista, hypo_Lista, syn_Lista) share common subset with
15:      (hyper_Listd, hypo_Listd, syn_Listd)){
16:      RETURN this.sensei;
17:      BREAK;}
18:    ELSE{GET hypernymi FROM hyper_Lista; //if above step do not work, go to
19:      FOREACH (hypernymi:hyper_Lista){ //higher level, use hypernyms instead of
20:         $w^a$  = hypernymi; // current word, repeat prior steps
21:        REPEAT Line 6-17; //if still not working, word is not disambiguated
22:        RETURN NULL;}}}} // RETURN NULL as result
23: //Show the result
24: PRINT ( $w^a, \text{sense}_i$ );
25: //END

```

Figure 4.5. The F-WSD Algorithm

We design the *F-WSD* algorithm based on following design rationale. In a term t , given a surrounding n -gram bag-of-words W^d , the target word w^a can be disambiguated if: i) w_i^d ($\exists w_i^d \in W^d$) appears in the hyponyms, hypernyms, or synonyms of w^a ; or ii) if hyponyms, hypernyms, or synonyms of w^a and w_i^d shares common subset(s); or iii) if former two conditions are not met, substitute w^a with one of its *direct* hypernyms instead, and repeat previous step. If none of the three conditions is met, the algorithm will return a null value indicating that no disambiguation suggestion can be provided based on given feature values. Essentially, the three conditions listed above can be recognized as classification rules within a logical sequence; thus, decision trees can be used to represent them, which are used to recursively partition the data set. In this context, the data set would be the words in the selected terms requiring disambiguation; the branches are the states in the disambiguation process, the nodes reflect aforementioned conditions, while the leaves are the senses (or *null* value if none sense is selected). The terms that could not be disambiguated using *F-WSD* could possibly be disambiguated via the integration of domain taxonomies (ontologies), by aligning the term with the corresponding term (and its *sole* sense) in the domain taxonomy/ontology.

Certain ground rules need to be established for the WSD process, rules similar to those found in the research article by Meijer et al. (2014). Normally, the terms that need disambiguation would appear in the text corpus several times; thus, by applying the *F-WSD* algorithm on them, multiple senses would be (*possibly*) assigned to a term – which is against the spirit of WSD. To solve this issue, we only allow *one* sense for each term: the most frequent sense is selected as the disambiguated sense of the term. On the other hand, a sense should correspond to one and only one term in the term collection extracted/filtered from the text corpus. If two different terms (i.e. t_1 and t_2) share the same sense after the WSD process, t_1 and t_2 are treated as synonyms ($t_1 = t_2$). Such synonym relations are useful for subsequent ontology integration as well as relation/property identification purposes.

4.3.3 Ontology Integration

The ontology integration process includes two sub-steps, namely term enrichment and seed concept expansion.

We enrich the selected and disambiguated terms with the synonyms/acronyms from the same domain (i.e. “*negative revenues*” and “*losses*”). Further, similar to the term enrichment approach reported in (Jiang & Tan, 2005; Kang et al., 2014), we design a mechanism in the light of enriching multi-word terms. The differences between our method and the methods in (X. Jiang & Tan, 2005; Y.-B. Kang et al., 2014) are: i) we use a post-selection enrichment, which would reduce the computation complexity of the term extraction and selection phase; and ii) we rely on the ranking of the term based on their DRs from the second part of Equation (1), and then enrich the nouns with rankings higher than a pre-defined threshold – rather than traversing through all the nouns in the selected terms. For instance, given a term ($\mathbf{NN}_1, \mathbf{NN}_2, \mathbf{VB}_1, \mathbf{VB}_2, \mathbf{NN}_3$), as well as a pre-defined threshold at 0.7 – if $\text{DR}(\mathbf{NN}_1) = 0.8$, $\text{DR}(\mathbf{NN}_2) = 0.9$, and $\text{DR}(\mathbf{NN}_3) = 0.6$; only \mathbf{NN}_1 and \mathbf{NN}_2 are chosen for enrichment.

The next step is to expand the seed concept list with the selected terms. There are several ways of updating existing ontology with newly discovered terms (Dorji et al., 2010; Gaeta et al., 2011). In this project, we proposed a semantic similarity based approach for such purpose. This approach relies on a similarity matrix, in which each cell represents the similarity between a newly discovered term t_n and an existing term t_e in the seed concept list. Semantic similarity has been widely applied in NLP and Information Retrieval domains, which is termed as a measure of semantic relatedness reflects the semantic relationship (such as “*is-a*” or “*a-kind-of*”) based on information theory (Resnik, 1995). A large number of measures with respect to semantic similarity has been published in the literature (Leacock & Chodorow, 1998; D. Lin, 1998; Resnik, 1995), which can be categorized as *corpora-based* and *knowledge-based* metrics – a detailed discussion of such categorizations can be found in (Sánchez, Batet, Valls, & Gibert, 2009). *Corpora-based* metrics rely on the co-occurrence of a pair of terms within the document corpus, while *knowledge-based* metrics map the terms representing concepts in a formal knowledge structure (such as WordNet or other domain ontologies). The *knowledge-based* measures are more preferable in this work since they rely on knowledge networks rather than *enormous* document corpus or *implicit* external knowledge (Ge &

Qiu, 2008). A disadvantage of *knowledge-based* approach is that if a term cannot be mapped to the knowledge structure, the measure of semantic similarity is impossible. However, this is not an issue in this project because i) we are updating ontologies – such terms can be treated as new classes in the existing ontology, and ii) the WSD phase reduces, if not eliminates, the possibilities of the “*lack-of-mapping*” issue. Among various *knowledge-based* semantic similarity metrics, we select the *WuP* measure rather than the others – this particular metric measures the normalized depth of concepts and their *Least Common Subsume* (LCS) (Wu & Palmer, 1994). The rationale behind such design decision is that the *WuP* measure relies on relative depth, and it is normalized, thus it enables orchestrations with extremely complex ontology (such as US-GAAP, WordNet, or the Gene Ontology), as well as enables the comparisons across different ontologies. The *WuP* measure is calculated as follows:

$$sim_{WuP}(w_1, w_2) = \frac{2 \times depth(LCS, root)}{depth(w_1, LCS) + depth(w_2, LCS) + 2 \times depth(LCS, root)} \quad (2)$$

In Equation (2), *LCS* refers to the farthest shared parent of a pair of words (w_1, w_2) according to the knowledge structure, whereas *root* denotes the root node in it (i.e. *Thing* in WordNet); *depth* is the number of intermediate nodes between two nodes. Figure 4.6 illustrates an example: $depth(w_1, LCS) = 3$, $depth(w_2, LCS) = 5$, while $depth(LCS, root) = 2$. Based on Equation (2), $sim_{WuP}(w_1, w_2) = \frac{2 \times 2}{3 + 5 + 2 \times 2} = 0.33$.

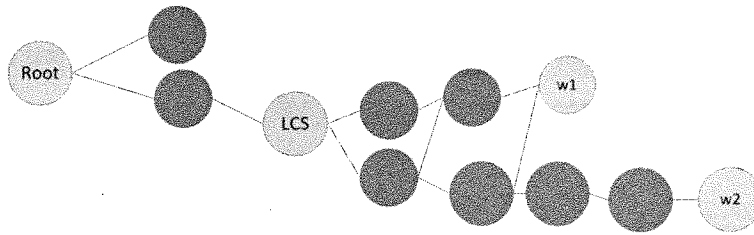


Figure 4.6. *WuP* Calculation Example

Moreover, the original *WuP* measure reflects the semantic relatedness of two single words. However, in order to align two terms, we need a measure to calculate similarity between multiple-word terms. Thus, we propose the *Normalized Multiple Word Semantic Similarity* (NMWSS) as follows (Equation 3), where $t_n = \{w_i | i = 1 \dots m\}$ and $t_e = \{w_j | j = 1 \dots n\}$ are two multiple-word terms, respectively. If the *NMWSS* between a

newly discovered term t_n and an existing term t_e is greater than a pre-defined threshold, then t_n is added to t_e as sub-class/instance; otherwise, a new (sibling) class needs to be created.

$$sim_{multi}(t_n, t_e) = \sqrt{\frac{\sum_{i=1}^m \sum_{j=1}^n (sim_{wup}(w_i, w_j))^2}{m+n}} \quad (3)$$

4.3.4. Semantic Relation Extraction

Abovementioned ontology integration step (discussed in Section 2.3) resolves the taxonomical – also known as hierarchical (Buitelaar, Cimiano, & Magnini, 2005) or vertical (Hwang, Choi, & Kim, 2011); however, a mature ontology enrichment approach needs to deal with semantic (non-taxonomical) relations between concepts – the learnt relations serve as the foundations for reasoning and other knowledge applications. Semantic relations are integral components in ontology learning and enrichment procedures; however, comparing to the abundance of studies on taxonomical relations, non-taxonomical relations have received less attentions (Punuru & Chen, 2011). Usually, there are three schools of relation learning approaches: statistical (i.e. machine learning based approaches (Li, Wang, & Khan, 2013; Zhang, 2008)), linguistic (i.e. syntactic pattern based approaches (Fader, Soderland, & Etzioni, 2011)), and hybrid approaches (Punuru & Chen, 2011; Shen et al., 2012). Recent efforts toward relation extraction have been focused on the hybrid approaches since they do not require a large, quality training set (i.e. well-annotated corpus), and they can better represent the semantics; thus, we decide to adopt the hybrid method in this study.

Before further discussion into the relation extraction, we first define the notion of *binary* semantic relations: $t_1 \rightarrow R \rightarrow t_2$, in which, $t_{1,2}$ refer to any two terms integrated in the existing taxonomy/ontology, while R refers to the collection of all possible *intra-sentence* relations between these two extracted terms. For capturing the lexico-syntactic patterns in text corpus, we adopt the well-accepted Subject-Verb-Object (SVO) tuples. A relation r ($r \in R$) is an instance of a variant of the SVO tuple; we propose three explicit variants/patterns and one hidden variant/pattern of the SVO tuple, which respectively are: $([S], [V], [O])$, $([S], [V])$, $([V], [O])$, and $([S], ?, [O])$. A table of the patterns and examples

can be found below (Table 4.2). Note that elements in these sets are not sequential (for instance, both (S, V, O) and (V, S, O) match the first pattern $([S], [V], [O])$); further, each element in a tuple can replicate more than once (for example, (S, V, O, O) also matches the first pattern) - which distinguishes our approach with the approaches proposed in (Punuru & Chen, 2011; Sánchez et al., 2012). For the three explicit patterns, the verb (VB) in them indicates the nature of the relation; while in the hidden pattern $([S], ?, [O])$, the labeling is more complicated.

Table 4.2. Selected lexico-syntactic patterns and examples

Pattern	Example
$([S], [V], [O])$... future losses in foreign markets will harm our financial health ...
$([S], [V])$..., where the losses incur ...
$([V], [O])$... thus, the possibility of future losses needs to be calculated ...
$([S], ?, [O])$... intellectual property infringement , a type of risks , would lead to ...

To better utilize relations extracted from text, a few properties need to be examined if a particular relation r holds, these properties include: *symmetry*, *reflectivity*, and *transitivity*. We define these properties as follows:

- **Symmetry**: given a relation r holds between two terms t_1 and t_2 ($t_1 \rightarrow r \rightarrow t_2$), if the same relation r holds when switching the two terms ($t_2 \rightarrow r \rightarrow t_1$), then r is symmetric. For instance, $r = \text{accompany with}$ stands when $t_1 = \text{losses}$ and $t_2 = \text{legal risks}$.
- **Reflectivity**: given a relation r holds between two terms t_1 and t_2 ($t_1 \rightarrow r \rightarrow t_2$), if the same relation r holds when $t_1 \rightarrow r \rightarrow t_1$, then r is reflective. However, as stated in (Sánchez et al., 2012), the reflectivity property is usually used between a term and its co-references (i.e. $t_1 \rightarrow r \rightarrow \text{itself}$).
- **Transitivity**: this property of a relation is checked in two steps: firstly, given a relation r holds between two terms t_1 and t_2 ($t_1 \rightarrow r \rightarrow t_2$), if the same relation r holds between t_2 and another term, t_3 ($t_2 \rightarrow r \rightarrow t_3$), then r is partially transitive; moreover, if the same relation r holds between t_1 and t_3 ($t_1 \rightarrow r \rightarrow t_3$), then r is transitive. For instance, the relation “*incurrence*” holds among terms “*legal disputes*”, “*slide in market share*”, and “*loss in revenues*”.

With the definition of the semantic relations, as well as their properties, we can move on to the extraction of relation among extracted terms from text corpus.

We propose a two-stage approach for extraction of semantic relations from domain-specific text corpus, namely *relation identification* and *relation selection*. Relation identification refers to the process of identifying instances matching pre-defined lexico-syntactic patterns in texts (relation candidates); while relation selection refers to the selection and labeling of discovered relation candidates. Figure 4.7 below depicts the overall workflow of the relation extraction process. The process depicted in Figure 4.7 is very straight-forward. After a sentence is located in a document, the grammar tree is extracted for the purpose of detecting clauses and dependencies. The dependencies are automatically selected as initial relation candidates, and compared with existing axioms from the domain. If a match is found, the dependency is selected as a non-taxonomical relation; if otherwise, the candidates are compared with patterns in Table 4.2 – *vf*icf* metrics for conformable candidates are computed and ranked. As a last step, top ranking candidates (with user defined threshold, i.e. top 50 candidates) are selected as relations and added to the *enriched* ontology.

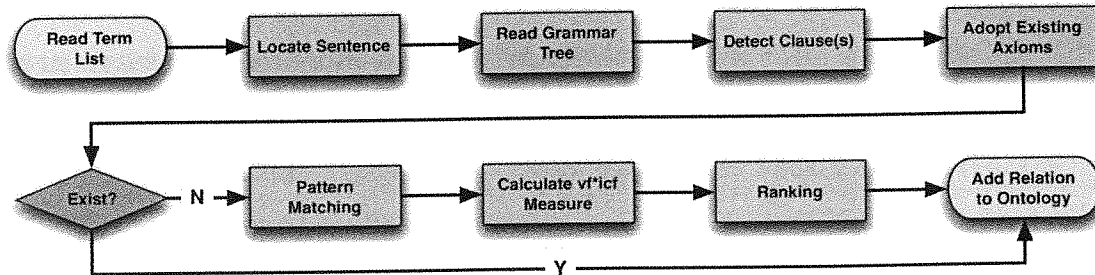


Figure 4.7. Flow chart of relation extraction

Unlike related prior studies (that extract relations at *sentence* level, i.e. Jiang et al. (2007), Sánchez et al. (2012), Shen et al. (2012), Weichselbraun, Wohlgenannt, & Scharl (2010), Zhang (2008)), our approach aims at extracting relations from a more granulated *clause* level. The term “clause” used here is somewhat different than the term in linguistics: in traditional linguistics, a clause, particularly a main clause needs a subject within its span; however, clauses from the context of computational linguistics can share the subject of the main sentence – meaning that a subject can be absent from the clause,

as long as it shares the subject of the main sentence. Essentially, we follow the pattern definitions of clauses from the discussions in a related study (Corro & Gemulla, 2013), which is adopted from the traditional linguistics to the computational linguistics domain. An illustrative example can be found in Section 5.1.4. The rationale behind such design decision is that during the experiments in our pilot study, we have found that extracting relations at the sentence level would return lower accuracies (60 – 70% on average), since the sentence structures will hinder the detection of the accurate semantics. For instance, a “*but*” structure would change the overall sense of a sentence – thus, it is reasonable to select clauses as the syntactic units for analysis.

The relation extraction process begins with reading in the term list, in which contains the extracted and filtered terms. The second step in the process is to locate the sentences containing the terms read in the prior step; only the sentences with two or more terms are selected. Then, the grammar trees (Jiang et al., 2007) of the selected sentences are analyzed, in order to identify the verbs in the selected sentences. As discussed earlier, those verbs are the indicators of semantic relationships. The verbs are also disambiguated and organized according to the WordNet taxonomy. In the next step, the existing axiom(s) between a pair of terms is extracted from according domain ontology, because the terms are already aligned to existing domain ontologies, there is a chance that axioms/relations already exist between them. For instance, given two terms “revenues” and “losses” extracted from prior steps, the ontology in the finance domain (i.e. *US GAAP*) is retrieved and based on it, an axiom “inversedGlossary” is identified between them; thus, a relation is discovered between the two terms (*revenues* \rightarrow *reversedValue* \rightarrow *losses*). Further, not only the nature but also the properties of the axiom is inherited (i.e. the relation “inversedGlossary” is symmetric). The inheritance of domain axioms will partially solve the identification of the hidden pattern ($(([S], ?, [O]))$). However, there are some term pairs that no axiom/relations are defined for them (which is the purpose of the enrichment as suggested in this work); hence, the process moves on to the pattern matching step, in which the core is the identification of the verb(s). The verb(s) within a clause defines the type of the clause – we adopt the types of clauses as suggested in (Corro & Gemulla, 2013). Identified by the verb(s), the clauses are used to match the syntactic patterns discussed in Table 4.2. Along

with prior related studies (Hwang et al., 2011; Ittoo & Bouma, 2013a; Maslennikov & Chua, 2010), we define a ranking metric following the spirit of *tf*idf*, namely *vf*icf* (*verb-frequency-inversed-clause-frequency*). In which, *verb frequency* refers to the occurrences of a specific verb in a text corpus; while *clause frequency* is the occurrence of clauses containing the verb of the same type (decided in prior step, a possible scenario is the same verb). The underlying logic behind the *vf*icf* metric is the *vf* part assures that the frequent mentioned verbs have higher possibilities to be selected, while the *icf* part is an assigning factor that assigns higher selective value to distinguished verbs. As in the next step, verbs are aggregated (into relations) by type and ranked based on the *vf*icf* metric. In the last step of the relation extraction process, relations with the aggregated *vf*icf* values higher than the predefined threshold are selected and added to the target ontology. Following equation aggregates the *vf*icf* metric of the verbs into relations (Equation (4)):

$$vf * icf_R = \sum \overline{sim}_v \times vf * icf_{verb} \quad (4)$$

In which, \overline{sim}_v is the *WuP* similarities among verbs, which serves as the weights in the aggregation. Considering the similarities among verbs assures the aggregated relation better reflects the semantics of the atomic verbs. The aggregation of verbs into relations will also partially benefit the identification of the hidden pattern ($([S], ?, [O])$) – given a certain pair of terms, even if the verb is missing, the relation with the highest *vf*icf* value between them could be used as a substitution. To the best of our knowledge, this is the first work that combines the syntactic patterns with machine learning approaches, utilizing the novel *vf*icf* metric; as well as the verb aggregation mechanism.

The proposed method is realized in a research prototype, which is described in Section 5.2.

4.4. Predictive Analytics Module

Results from previous modules provide required data for predictive modeling targeting IPO pricing trends. Different modeling techniques are selected from pre-defined repository. Within this module, two major sub-modules need to be highlighted, namely

Model Construction and Result Validation & Communication. Detailed discussions toward each sub-module can be found below.

4.4.1 Model Building

Based on the analytical problem (IPO Pricing Predication), two phases are required in the Model Construction sub-module, which are variable definition/selection and modeling technique selection/operationalization. Variable definition/selection refers to the process of defining/selecting predictor(s) and target variable(s) for predictive models. Modeling technique selection requires analysts to select/operationalize proper predictive modeling techniques for given analytical problem and variables. The details of this sub-module are discussed below.

A. Target Variable Definition

The target variables in this study include pre-IPO price revisions (*PRCREV*) and post-IPO price trends (*X1stDay*). We limit the scope of post-IPO price trends to the first trading day, meaning the trends are obtained by comparing the first day opening price (final offering price, P_{ipo}) and the first day closing price (P_{1day}). In order to define both target variables, we first define the pre-IPO price adjustment (ΔP) and the post-IPO initial returns (IR) defined as follows:

$$\Delta P = \frac{P_{ipo} - P_{mid}}{P_{mid}}, IR = \frac{P_{1day} - P_{ipo}}{P_{ipo}} \quad (5)$$

P_{mid} , P_{ipo} , and P_{1day} are the mid-point of the initial offering price range, the final offering price in the 424B4 filings, and the first-day trading price, respectively. The initial offering price range is obtained from the S-1 filings.

The target variables are binary variables, which are created based on the values of ΔP and IR in our sample, which are defined as follows:

- *Pre-IPO Price Revision (PRCREV)*: set to 0 if the final offering price is lower than or equal to the midpoint of the initial offering price range ($P_{ipo} \leq P_{mid}$), set to 1 if the final offering price is greater than the midpoint of the initial offering price range ($P_{ipo} > P_{mid}$);

- *Post-IPO First Day Return Change (X1stDay)*: set to 0 if the final offering price is lower than or equal to the midpoint of the initial offering price range ($P_{1Day} \leq P_{ipo}$), set to 1 if the final offering price is greater than the midpoint of the initial offering price range ($P_{1Day} > P_{ipo}$).

B. Predictor Definitions

In this study, we have three sets of predictors. The first two sets correspond with aforementioned two critical sections in IPO prospectus, namely “Risk Factors” and “Management Discussions and Analysis”. Thus, we define these predictors as *risk-features* and *MD&A features* from now on. The third set of predictors reflects important financial and other characteristics of IPOs. Thus, we define them as *predictors from prior studies*.

The *MD&A features* and *risk-features* are defined as follows. In the *MD&A* sections, for each 424B4 filing for an IPO, a total of 15 counts are returned (5 features x 3 sentiments). It may be noted that a sentence might overlap over different features/sentiments – for instance, a sentence might discuss both *sales* and *net income*, or contains a transition structure where the first clause is denoting a positive sentiment while the second one is negative. The counts are then normalized to account for variation in lengths of different *MD&A* sections. In the *Risk Factors* sections, a total of 6 counts (each one corresponding to a unique risk-feature) are returned for each Form 424B4 filing. Note that sentences might overlap over different features – similar to the *MD&A* sections. In addition, hereby we introduce two new predictors (mediators), namely *MDAScore* and *RiskScore*. These two are aggregated predictors for pre- and post-IPO prices. Details regarding the calculations of these two variables are discussed in Equation (6) and (7) on Page 48 of this dissertation. Brief definitions of these predictors are shown in Table 4.3 below.

Table 4.3. Predictors Used in Predictive Models

Variable Name	Description
Predictors	
Risk Features	6 features created by domain experts for significant business and managerial aspects discussed in the “Risk Factors” sections in the prospectuses
MD&A Features	15 features created by domain experts for significant business and managerial aspects, along with sentiments discussed in the MD&A sections in the prospectuses
RiskScore	Aggregated predictor from risk features; computed using linear relations between risk features and RiskWeight
MDAScore	Mediator, aggregated predictor from MD&A features; computed using linear relations between MD&A features and MDAWeight
Predictors from Prior Studies	
Up Revision	Set to $ \Delta P $ if $\Delta P > 0$, otherwise 0
Days between S-1 and 424B4	The logarithm of calendar days between the initial S-1 filing and the filing of Form 424B4 from EDGAR.
Top-tier Dummy	Dummy variable, set to 1 if the leading underwriter of IPO has a rating of 8 or higher, otherwise 0
Positive EPS Dummy	Dummy variable, set to 1 if trailing earnings per share is positive at the time of IPO, otherwise 0
Share Overhang	The ratio of retained shares divided by the number of shares in the IPO
Sales	Trailing annual sales/revenues in thousands of dollars at the time of IPO
Prior NASDAQ 15-day returns	The buy-and-hold returns of the CRSP NASDAQ value-weighted index of the 15 trading day period prior to the IPO date, ending on day $t-1$.

Besides aforementioned two sets of predictors, a set of predictors commonly used in prior related studies, which describe the IPO characteristics, are included in the predictive models. These predictors are selected from the finance literature of similar studies (Hanley & Hoberg, 2010; Loughran & McDonald, 2013). According to the literature, they have substantial impacts on IPO pricing from different perspectives. The descriptions toward this set of predictors are presented in Table 4.3 above.

In addition to aforementioned experiments, we are interested in looking into the prediction powers of the informative contents in the respective two sections (*Risk Factors* and *MD&A*), regardless of features and sentiments, toward pre- and post-IPO pricing trends. In other words, it is imperative to discover the link between management's

awareness of risks and confidence of future performances. Thus, in consistent with (Hanley & Hoberg, 2010), we designed a two-stage experiment for this purpose. We run the first-stage regression with *MDAWeight* and *RiskWeight* as the dependent variable, while the 15 sentiment-feature variables (*Sale*Positive* through *Cash*Uncertain*) and 6 risk-features as independent variables, respectively. The two weights (*MDAWeight* and *RiskWeight*) are the ratios of meaningful sentences against total sentences in respective sections. The purpose of this regression is to identify the weight of each feature-sentiment variable, and use them to calculate a sentiment-sensitive score for the *MD&A* and *Risk Factors* section of each filing – thus, the intercept is excluded from the model.

$$MDAWeight = \sum_{f=1}^5 \sum_{s=1}^3 \omega_{f,s} \times p_{f,s} + \varepsilon \quad (6)$$

$$RiskWeight = \sum_{f=1}^6 \omega_f \times p_f \quad (7)$$

The model is shown as Equation (6) and (7). For the *MD&A* section, $p_{f,s}$ is any of the 15 feature-sentiment variables (i.e. *Cost*Negative*), while the parameter estimates $\omega_{f,s}$ is the corresponding weight of the feature-sentiment variable.). For the *Risk Factors* section, p_f is any one of the 6 risk-features (i.e. *customers*), while the parameter estimates ω_f is the corresponding weight of the risk-feature. With normalized weights for each of atomic, we can calculate a weighted score (*MDAScore* or *RiskScore*) for the *MD&A* or *Risk Factors* section of each filing. Then *MDAScore* and *RiskScore* are used in the second-phase predictive models as predictors.

4.4.2. Modeling Techniques

In this study, six distinguished predictive modeling techniques are selected for predicting IPO pricing trends – since the target variables in the predictive models are binary, we select widely used modeling techniques such as *Logistic Regression* (LR), *Decision Tree* (DT), *Artificial Neural Networks* (ANN), *Support Vector Machines* (SVM), *Random Forest* (RF), *k-Nearest Neighbors* (k-NN), and *Ensemble* (EN). The selected predictive modeling techniques are implemented in the statistics toolkit SAS Enterprise Miner (Version 13.1). Brief descriptions of these modeling techniques are provided as follows (Hair et al., 2010):

- Logistic Regression (LR): with binary target variables (*PRCREV* and *X1stDay*), logistic regression is used to estimate the classification (usually either 0 or 1) by the maximum likelihood – the target variable is estimated through a linear function of the logarithms of the predictors. For comprehensive coverage, we also included General Linear Modeling (GLM), which is similar to LR – relying on linear relations.
- Decision Tree (DT): decision tree is another data analysis tool that uses a tree-like model for decision support purposes; in predictive analysis, decision trees are used to map observations of an event to the target value. Leaves in decision trees correspond to class labels, while branches reflect conjunctions of features toward these class labels. Advantage of DT includes its simplicity to interpret, white box model, and robustness; while its disadvantage is mainly with the optimization. In this study, DT is an appropriate predictive modeling technique due to the nature of the decision problem and the characteristics of the data.
- Artificial Neural Network (ANN): ANN usually refers to the set of machine learning techniques that is inspired by the biological neural network (i.e. the brain), and is used to estimate the unknown value(s) of the target variables. Atomic analytical units in ANN are refereed as neurons, which conduct machine learning tasks by computing values from input data. However, the computational complexity of ANN is high when more layers exist in the predictive model.
- Support Vector Machines (SVM): Unlike LR, SVM is defined as a non-probabilistic binary linear classifier. SVM is supervised machine learning models, while relies highly on the training data set. The data points in the training data set are mapped into a space, and categorized with a gap; while testing data are mapped into the same space and predict to aforementioned categories.
- Random Forest (RF): by the name, RF grows many DTs – the logic is the input data is in the form of a vector along with the trees in the forest, in order to classify a new object. The final results are chosen with most votes from the trees. RF is good at generating unbiased estimate of the generalized error, and reflecting the relation between each predictor and the classification. However, it is not as good as other predictive modeling techniques with respect to accuracy.

- **k-Nearest Neighbors (k-NN):** k-NN is another major classification algorithm, which is one of the algorithms with lowest computational complexity. In k-NN, k is a user defined constant (a positive integer). An item is assigned to a cluster by the voting of adjacent clusters – within the space created by vectors from the training sample. The item is assigned to a cluster that is most common among k neighbors of this cluster, according to a distance measure. In this study, we use Memory Based Reasoning as the k-NN classifier.
- **Ensemble (EN):** ensemble is a machine learning method that multiple models are trained and used to solve the same decision problem. Ensemble learning is implemented due to two advantages, namely accuracy and efficiency. Ensemble improve prediction accuracy since uncorrelated errors of individual modeling techniques can be minimized via ensemble averaging; while efficiency refers to that in some cases, the prediction is not available with single models, but maybe approximate by ensemble averaging.

The selected modeling techniques used in the experiments, along with their abbreviations, are shown in Table 4.4 below.

Table 4.4. Modeling Techniques Used in the Predictive Modeling Module

Model Number	Modeling Techniques	Abbreviation
1	Classification And Regression Tree – Decision Tree	DT
2	Random Forest	RF
3	Ensemble	EN
4	Artificial Neural Network	ANN
5	Logistic Regression	LR
6	Generic Linear Model	GLM
7	Support Vector Machine	SVM
8	Memory Based Reasoning – k-Nearest Neighbor Clustering	k-NN

With selected predictive modeling techniques discussed, in the following subsection the model validation and selection strategies are discussed as well.

4.4.3 Model Validation and Selection Strategy

For validating aforementioned predictive models, two issues need to be dealt with, namely *overfitting* and *imbalance*.

Overfitting is a popular issue in the domain of machine learning, which occurs when instead of reflecting the underlying relationship to the target function, the predictive model depicts the noise and has poor predictive power (Zhong, Li, & Wu, 2012). Reasons of overfitted model is that the model is unnecessarily complex – i.e. contains too many variables, or the model structure is not conformable with the data. A practical explanation of overfitting is the model is trained and fit to the training data (known data), and not as generalizable to the testing data (unknown data). Dimension reduction is a well-adopted approach to deal with overfitting issue in machine learning models. Different dimension reduction methods are associated with different predictive modeling techniques, such as *tree pruning* for DT, *support vector reduction* for SVM (Ngo-Ye & Sinha, 2014); as well as other generic approaches such as *Principle Component Analysis* (PCA) used in (Song, Yang, Siadat, & Pechenizkiy, 2013). In this study, appropriate dimension reduction methods (such as tree pruning and PCA) are adopted – details are provided in Section 5.4.

Another issue has to do with imbalanced data, particularly with the target variables, which is very significant in this study. From previous relevant studies (Firth, Wang, & Sonia, 2013; M.-C. Lin et al., 2011), stock prices tend to increase, particularly during the first trading day (*X1stDay*). Our data corresponds with such observation – in the selected sample (as reported in Table 4.2), 414 out of 517 IPOs (80.08%) have the first day initial returns are greater than the offering prices ($IR > 0$), while only 19.92% have negative IRs. As a comparison, the ratio between positive and negative ΔP is 61.12% to 38.88%. Highly imbalanced data will impair the prediction accuracy of the models. There are several approaches to deal with imbalanced sample, such as increasing the sample size, oversampling the minority class, undersampling the majority class, or combining the prior three in a systematic manner (Wasikowski & Chen, 2010). The resulting data set ends up with balanced data set. In this study, we select the oversampling method – which increase the data points in the minority ($IR > 0$) class. It is worth noting that the oversampling treatment only affects the training data set, not the validation and testing sets, to keep the prior probabilistic. The rationale behind the design decision is that even though oversampling may lead to overfitting issues – which can be

mitigated using aforementioned techniques – it is still the best imbalance handling approach, given the limited sample size in the selected data set.

Another essential data preprocessing step is data partition. Data partition is an efficient strategy to validating the quality and generalizability of (predictive) models. A (large) portion of the selected data set is used for fitting the predictive models preliminarily – this portion of the data is named as the training data set. The remainder of the data set serves for the empirical validation purposes toward the model trained using the training data set. Normally, the remaining data is further partitioned into two groups: validation data set and testing data set. The validation data set is an efficient approach of preventing models from overfitting (toward the training data), as well as comparing the accuracy and efficiency across different models; while the testing data set is used to ultimately assess the predictive capabilities of the models. In this study, we partition our data to training, validation, and testing data sets with the ratio of 65%, 20%, and 15% -- thus, our training, validation, and testing data sets end up with 336, 104, and 77 data points.

As discussed above, the validation and testing data sets are use to assess the predictive models. Thus, validation metrics need to be set up for such purpose. In this study, two sets of metrics are used in the validation step. The first set of metrics measures the accuracy of the predictions, which are calculated using the contingency tables discussed above. These metrics include precision, recall, accuracy, and F-score. They are calculated using following formulas (8a – 8d):

$$precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (8a)$$

$$recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (8b)$$

$$accuracy = \frac{True\ Positive + True\ Negative}{Positive + Positive} \quad (8c)$$

$$F - score = 2 \times \frac{precision \times recall}{precision + recall} \quad (8d)$$

Usually, precision, recall, and accuracy are usually not used in isolation. We determine to use the F-score as a single value for model selection purpose.

The other set of metrics depicts the performance of the predictive models, respectively. In this category, we select *Cumulative Lift* (defined as Lift from now on) and *Area Under Curve* (AUC, calculated from *Receiver Operating Characteristics*, ROC) to evaluate the model effectiveness. Lift and AUC are used to interpret the results of the predictions, which is consistent with prior related studies (Khansa & Liginlal, 2011; Wasikowski & Chen, 2010).

4.4.4. Result Validation and Communication

Validating results by the means of accuracy and performance is a very important step in data analysis (A. Sen, Dacin, & Pattichis, 2006). It includes several steps, including data filtering, result validation, and gross error detection, in terms of accuracy. All these validation techniques are adopted in validating and communicating the results. For instance, standardization procedures are employed for all predictors, as a data-filtering step. Also, different sets of evaluation metrics toward the predictive results are employed in order to test the accuracy and performance of the prediction.

On the other hand, communicating predictive results to domain experts ensures the relevance of the study, as discussed in Section 3.1. With the help of data visualization and interpretation techniques, experts from the finance domain are able to understand and utilize the predictive results for investment decision support, and other practical purposes. Details regarding this sub-module can be found in Chapter 5, along with the discussion of the results.

CHAPTER 5

DEMONSTRATION AND EVALUATION

In this chapter, we firstly present the demonstration and evaluation of the Ontology Enrichment module – the demonstration is conducted via a standalone case study. Secondly, the workflow of the predictive modeling, which illustrates the relevance of this work, is demonstrated and discussed in Section 5.2. Section 5.3 discusses the details regarding the data used in this study, as well as the descriptive statistics. Section 5.4 presents and discusses the results from the predictive models, with respect to the context of IPO pricing prediction. The Information Extraction module is demonstrated in Section 4.2 in previous chapter, and evaluated along with the two modules mentioned below.

5.1 Ontology Enrichment Module

The pipeline realizing the Ontology Enrichment module is depicted in Figure 5.1. In order to deliver a flexible and extensible system, we have adopted *GATE* as the orchestration mechanism for our system (Cunningham et al., 2002). *GATE* is a widely applied NLP toolkit based on Java-like rules (*JAPE*, *Java Annotations Patterns Engine*), developed for IE and other analytical purposes. *GATE* provides a variety of packaged analytical/processing functionalities (namely Processing Resources, PR), such as *Tokenizer*, *Sentence Splitter*, and *NP Chunker*, for parsing the document corpus. *GATE* also allows users to encode other functionalities as *JAPE* rules (essentially a pattern-matching *left-hand-side* (LHS) and a Java program as a *right-hand-side* (RHS)), which are executed along with pre-built PRs in a pipeline-like fashion. The corpus itself, along with the domain ontology and WordNet, serves as Language Resources (LRs) in *GATE*. Moreover, other than running standalone, *GATE* can be embedded in other information systems (through provided Application Programming Interfaces, APIs) – so that other elements in the system, such as the User Dashboard, and an independent rule engine for querying/reasoning based on the ontology, can be developed. We implement all four major

modules, namely *Term Extraction*, *Term Selection*, *Ontology Integration*, and *Relation Extraction*, composing the proposed system as *JAPE* rules. Following sub-sections discuss the functionalities of different components within this module accordingly.

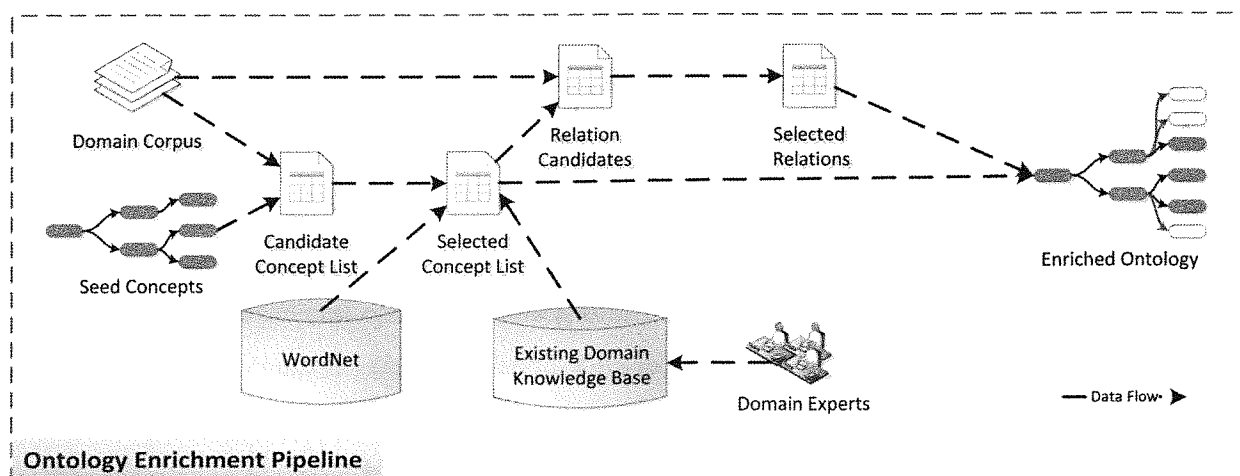


Figure 5.1. Ontology Enrichment Pipeline

5.1.1. Term Extraction Component

The *Term Extraction* module undertakes two major functionalities, namely *preprocessing* and *NP extraction*. In the preprocessing of the documents in the corpus, we apply a pre-built plugin in GATE, named *OpenNLP* (Open Natural Language Processing) for parsing the documents. *OpenNLP* is a native plugin incorporated in GATE, originally a library based on *Apache OpenNLP* library (Apache-OpenNLP-Development-Community, 2014). *OpenNLP* also follows a pipeline-like fashion. To make it match the purpose of our study, we updated the pre-built package by modifying the code and/or adding new PRs to it, which are discussed in detail below. The major components in the preprocessing sub-module include the following:

- *OpenNLP Tokenizer*: the *OpenNLP Tokenizer* splits documents into small tokens, such as words, numbers, punctuations, symbols, and spaces.
- *OpenNLP Sentence Splitter*: rather than the default sentence splitter, we select the *OpenNLP Sentence Splitter* in our pipeline and modified the original code to support further segmenting sentences into sub-sentences and/or clauses, which helps us in extracting the syntactic features for term selection purposes.
- *OpenNLP POS Tagger*: the *OpenNLP POS Tagger* assigns POS tags to tokens such as words and symbols with the default lexicon and rule sets. Moreover, the *OpenNLP Name Entity*

Recognition (NER) PR is incorporated in the pipeline, in order to annotate original *Message Understanding Conference* (MUC) entities, such as *person*, *location*, *organization*, *date*, and so forth. Such annotation is helpful in the later deep cleansing step.

- *Stemmer* and *Morphological Analyzer*: we adopt the two components for lemmatizing the tokens in the document corpus. After this step, all the morphemes (affixes, POS variants, etc.) of the same stem (root words) are annotated with additional features “*stem*” and “*root*” in the token annotations. For instance, a stem feature of “*convert*” is added to both tokens “*converting*” and “*convertible*”.
- *GATE* provides several options in order to implement the *NP Extraction* sub-module, such as *noun phrase chunker (NPChunker)*, *Tagger Framework*, *LingPipe NER PR*, and the *OpenNLP Chunker*. We select the *OpenNLP Chunker* in implementing the proposed system because: i) as a native PR in the *GATE OpenNLP* plugin, *OpenNLP Chunker* collaborates better with other *OpenNLP* components (such as the ones used in the preprocessing step); 2) as evaluated in a recent study (N. Kang, van Mulligen, & Kors, 2011), the *OpenNLP Chunker* yields in the highest *accuracy* and *ease-of-use* compared to its counterparts. The *OpenNLP Chunker* is essentially a JAPE rule – while a rule set is called in the LHS for linguistic pattern matching purposes. We modified the rule set by incorporating the patterns in Table 4.1 in the LHS to make sure it is capable of matching those linguistic patterns, while redundant patterns are removed. *OpenNLP Chunker* adds a feature to the tokens in the document, which uses the common BIO values: for instance, a token tagged with a “*B-NP*” value means that it is at the beginning of a noun phrase; while a token tagged with “*I-NP*” means it is inside a noun phrase. This feature is critically useful for identifying the local features as discussed in Section 4.3.1.

5.1.2 Term Selection Component

There are two phases in the *Term Selection Module*, namely *related term ranking* and *domain specific deep cleansing*. Several JAPE rules are encoded for implementing this module. Before calculating the ranking of the term candidates, a linguistic filtering needs to be conducted on them. The first group of JAPE rules is used for such purpose. The first JAPE rule is named “*StopWord-Remove*”, which removes the stop words from the extracted term candidates. The English stop word list is obtained from (Snowball-Tartarus, 2013). Then a JAPE rule named

“*Filtering*” is added to the pipeline – it has two main functions: filtering tokens with POS tags other than **NN**, **VB**, or **JJ**; and creating the *n*-gram bag-of-words based on the filtered words in the term candidates.

With the term candidates filtered, we can begin to calculate their ranking metrics. The first JAPE rule in this group is named “*Freq-Calculation*”, which calculates the term frequency according to the *FREQ* part of Equation (1). The second JAPE rule “*DR-Calculation*” computes the DR measure, according to the second part of Equation (1). A third JAPE rule “*Ranking*” calculates the final ranking measure, according to Equation (1), and then sort the candidate terms based on the calculated $rank(t, d)$.

The next chunk of JAPE rules conducts ‘*deep cleansing*’ on the words in the sorted term list. The LHS of the JAPE rule “*Deep-Cleansing*” matches the unwanted patterns based on the features from the tokens, while the RHS add a “DROPPED” feature to the corresponding token. A list of exemplar unwanted patterns from the case study can be found in Table 5.1.

Table 5.1. Rules Used in Deep Cleansing

No.	Rule //Explanation
1	<code>(Token.category = "NN" && (Token.kind = "LOC" Token.kind = "ORG")) //nouns of geographic locations, organizations</code>
2	<code>(Token.category = "VB" && Token.chunk = "O") //verbs outside any phrases</code>
3	<code>{(Token.category = "CD")}{(Token.category = "NN")} //nouns following a number</code>

As discussed in Section 4.3.1, the selection upon the term candidates are completed; and then the selected terms are used as input for the next module.

5.1.3 Ontology Integration Component

Two major phases exist in the *Ontology Enrichment Component*, which are *word sense disambiguation* and *ontology integration*.

To implement the WSD function in GATE environment, we employed a third-party plugin named *WordNet_Suggester* (Gooch, 2013). In essence, *WordNet_Suggester* is a pre-built JAVA toolkit that provides glossaries, hypernyms, hyponyms, synonyms, and other taxonomic relationships for a specific word, relying on WordNet (which loaded in GATE as a LR).

WordNet_Suggester provides us a gateway to retrieve the taxonomic information from WordNet, and it allows configuration through initialization parameters (such as *attemptFullMatch*: set to true if intend to match multiple words). However, we have to code a custom JAPE rule in order to realize the WSD method proposed in Section 4.3.2 (*F-WSD*). This rule uses output annotation set from the *WordNet_Suggester* of both the selected terms (outcomes of the *Term Selection Module*) and the *n-gram bag-of-words* surrounding the target word as inputs (LHS patterns), and implements the *F-WSD* algorithm on the RHS. It adds a feature “*WN_sense*” to the target word (token): if a sense is determined, then the sense is added as the value of “*WN_sense*”; otherwise, a *null* value is added.

The second phase in the *Ontology Enrichment Module* is ontology integration. For calculating the semantic similarity proposed above, we adopt *ws4j* (WordNet Similarity for Java) package (Shima, 2013) by calling its Application Programming Interface (API) in the JAPE rule “*simCal*”, which is used to calculate the *NMWSS* (as presented in Equation (3)) between words in two terms (discovered-existing pairs). Another JAPE rule, “*OntoSuggester*”, is developed to suggest the expansion of the seed concept list based on the calculation results from “*simCal*” and a user-defined threshold (as a runtime parameter). It is by design that the “*OntoSuggester*” rule does not directly update the concept list; instead it provides suggestions to the users/knowledge workers – in other words, it assists the concept expansion process, rather than replacing human judgments. Then variations of “*simCal*” and “*OntoSuggester*” are executed in the pipeline, between the pair of terms from the expanded concept list and the target ontology (the ontology requiring alignment).

5.1.4 Relation Extraction Component

As discussed in Section 4.3.4, there are two phases in the Relation Extraction module, namely *Relation Identification* and *Relation Selection*. We have encoded several JAPE rules in realizing them. For the purpose of illustrating the functionalities of this module, we select an example sentence from an IPO prospectus: “*Global capital and credit market issues could negatively affect our liquidity, increase our costs of borrowing and disrupt the operations of our suppliers and customers.*” The LHS of first JAPE rule “*SentenceLoc*” reads in the annotations of selected terms in the documents; the RHS does two jobs: locating the sentences with more than one terms from the selected term list, and reading the POS tags in the annotated sentences. The

retrieved POS tags are used to formulate the grammar tree of a certain sentence. The grammar tree (hierarchical structure) in the example sentence is shown in Figure 5.2. The second JAPE rule “*ClauseIE*” determines the spans of the clause, based on the verb(s) in the given sentence. In the example sentence, there are three verbs – thus, there are three clauses in the sentence (which is shown in Figure 5.2). The third JAPE rule in this module, “*ClausePM*”, determines the syntactic pattern of a specific clause, and then compares it with the pre-defined patterns.

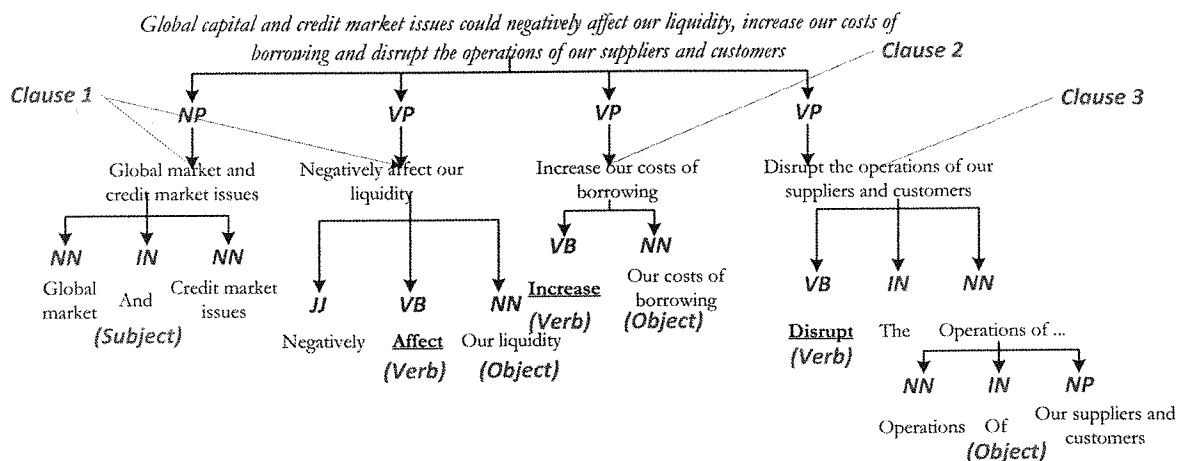


Figure 5.2. Grammar Tree and Clause Structure of the Example Sentence

As an example, the three clauses (shown in Figure 5.2) match pattern ([S], [V], [O]), ([V], [O]), and ([V], [O]), respectively. It is worth noting that even though clause 2 and clause 3 do not have a specific subject; the main subject in the sentence (“global market and credit market issues”) is reused as subjects in them as well – such definition is consistent with similar studies from information extraction and computational linguistics domain (Corro & Gemulla, 2013), as discussed in Section 4.3.4. The matched patterns (relation candidates) are stored in a list. The next JAPE rule, “*Vfief*” reads the stored relation candidates throughout the corpus, calculate the $vf*icf$ values based on Equation (3), and rank them. Only the relation candidates with the $vf*icf$ values greater than a threshold is added to the target ontology by the last JAPE rule “*RLPopulate*”. For instance, given the selection threshold as 0.6, if the $vf*icf$ values for the three relation candidates in Figure 5.2 are 0.45, 0.62, and 0.72; only the latter two are added to the target ontology.

5.1.5 Evaluation of the Ontology Enrichment Module: A Preliminary Case Study

The proposed system is instantiated in a research prototype '*IPO-Extractor*' and evaluated in a case study in the finance domain. The *IPO-Extractor* prototype is deployed on a machine equipped with an Intel Xeon 2.47 GHz CPU, 24GB RAM, and Windows 7 Enterprise 64-bit operating system.

The details regarding the case study are elaborated in following subsections, with some preliminary results and discussions.

A. Design of the Case Study

The case study aims to learn the knowledge structure regarding the IPO process through the textual contents in the IPO prospectus. Two domain experts created a seed concept list named the *IPO-Ontology*, which contains the key concepts with respect to the *Risk Factors* section in the prospectus. IPO prospectus is recognized as the most credible source when analyzing the phenomena within an organization's IPO process; whereas the *Risk Factors* section is deemed as one of the most information-rich sections within the prospectus (Hanley & Hoberg, 2010). It is well accepted that the textual information in the *Risk Factors* section has a significant effect on the IPO pricing volatility; yet no formal knowledge structure exists in the domain to support the IE-based analysis on it (Hanley & Hoberg, 2010; Loughran & McDonald, 2013). We plan to apply the *IPO-Extractor* on document corpus containing the *Risk Factor* sections of the IPO prospectus, for the purpose of enriching the *IPO-Ontology* for further analyses.

IPO-Ontology was developed with 6 first-level ontological classes, and 47 second-level classes in a hierarchical manner. The root concept is "*risk_factors*", and the first level classes include: *growth* (concepts related to the growth and business operations of an organization), *management skills* (the management views and strategies of a company), *competitiveness* (the ability to compete with the competitors), *customers* (the relation with current and potential customers), *lawsuit* (the capability to issue and react to a lawsuit, or potential lawsuit), and *stock prices* (the pricing strategy of the stock), which are the key factors disclosed in the *Risk Factor* sections that affect the IPO pricing. These factors are generalized from an intensive literature review in the finance domain. A snippet of *IPO-Ontology* (in OWL format) is shown in Figure 5.3.

In order to obtain the document corpus for the case study, we developed a web crawler to retrieve 424B documents from the EDGAR database of SEC. 424B documents are the final

prospectus in the IPO process. Several filtering rules created by the domain experts were applied in the web crawler, such as the rule that the company should not be in the finance industry (i.e. banking or insurance companies), the IPO should be within the period between 2003-1-1 to 2013-12-31, and common stock only. Any prospectus without a valid *Risk Factors* section is expunged. More than 2,000 424B documents were retrieved. A random sampling is conducted and a total of 400 documents were selected for this case study.

```

<owl:Class rdf:ID="competitiveness">
  <altMatch rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >protections</altMatch>
  <altMatch rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >protection</altMatch>
  <altMatch rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >competitors</altMatch>
  <altMatch rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >competitor</altMatch>
  <altMatch rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >compete</altMatch>
  <rdfs:subClassOf>
    <owl:Class rdf:ID="factors_risk"/>
  </rdfs:subClassOf>
  <prefMatch rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >competitiveness</prefMatch>
</owl:Class>
<owl:Class rdf:ID="management_skill">
  <altMatch rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >recruiting</altMatch>
  <altMatch rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >assimilates</altMatch>
  <altMatch rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >attracts</altMatch>
  <altMatch rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >personel</altMatch>
  <altMatch rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >executives</altMatch>
  <altMatch rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >executive</altMatch>
  <altMatch rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >personels</altMatch>
  <altMatch rdf:datatype="http://www.w3.org/2001/XMLSchema#string"

```

Figure 5.3. A Snippet of IPO-Ontology

Specifically for the purpose of the case study, a parsing/pruning step is conducted before applying *IPO-Extractor* on the document corpus. The parsing step aims at removing all non-textual contents in documents (including tables, figures, table of contents, file head, etc.) and annotating the *Risk Factors* sections from the 424B documents. The average length (number of word tokens) of the selected documents is 86,332; whereas the average length of the *Risk Factors* sections is 3,791 (4.39% of the average document length).

B. Results and Discussions

In this case study, we have selected the first 50 term candidates in the sorted list (using Equation (1)) and set the threshold for NMWSS to 0.5; also, we set the threshold for relation selection as 0.5 as well.

To evaluate the competitiveness of our method against other extraction methods, we have selected several *GATE*-based term extraction methods, including: *ANNIE* (A Nearly-New Information Extraction System) + *NPChunker*, *ANNIE* + *KEA* (Key Phrase Extractor). The same document corpus and thresholds are used for both other methods. For relation extraction purposes, we have selected *OLLIE* (Open Language Learning for Information Extraction) (Schmitz, Bart, Soderland, & Etzioni, 2012) as a complement to the *GATE* native method (using embedded *GATE*). We select *OLLIE* because it is a hybrid relation extraction tool (similar to our relation extraction approach) and it is more comprehensive with respect to relation extraction (Corro & Gemulla, 2013). We use the duration of the extraction process (*Duration*, in minutes) and the RAM usage at peak during the extraction process (*RAMUSE*, in Gigabytes) as indicators of the *efficiency* of different methods. The results are shown in Figure 5.4. From Figure 5.4, it is clear that *Duration* and *RAMUSE* of IPO-Extractor is evidently lower than the other two methods. These two metrics suggest that the proposed IPO-Extractor is more efficient than current *GATE*-based extraction methods. The reason for the performance improvement is because of: a) we take the context of the term(s) into consideration – according to Wimalasuriya & Dou (2010), taking the context into consideration will improve the performance of OBIE systems, as well as associated processes -- i.e., ontology enrichment (Meijer et al., 2014); b) the deep cleansing step in our approach, before the term selection, reducing the computational complexity of the ontology enrichment process (Wong, Liu, & Bennamoun, 2009).

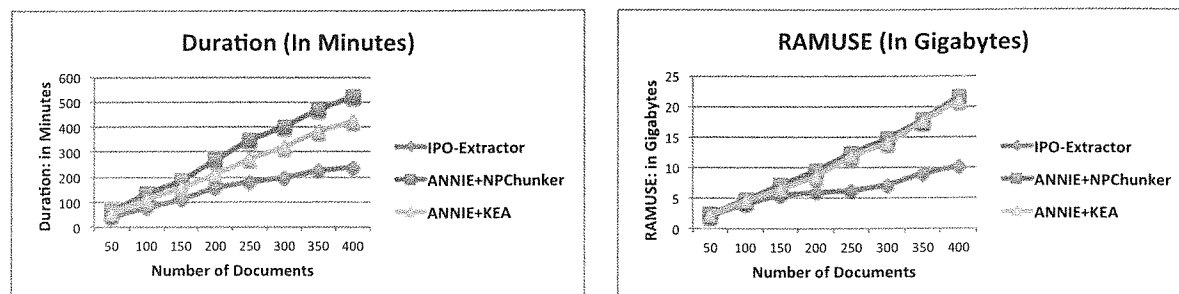


Figure 5.4. Efficiency Comparison Between Different Methods

For the purpose of evaluating our approach, three domain experts were asked to manually extract terms and relations from 200 out of the 400 selected documents. A total of 57 terms and

82 relations were extracted from the document corpus (with an inter-rater agreement of 87.8%, and with an agreement score of 0.73). Such extractions are used in the evaluation process as the “ground truth”.

We propose two sets of metrics for evaluating the quality of the proposed method. The first set has to do with the *coverage* extracted terms/relations. Coverage refers to how many correct terms/relations are extracted from the text corpus. The coverage of terms/relations of the three selected approaches is shown in Figure 5.5 below. It is obvious that the *IPO-Extractor* based on the proposed approach outperforms the two *GATE*-native approaches plus OLLIE.

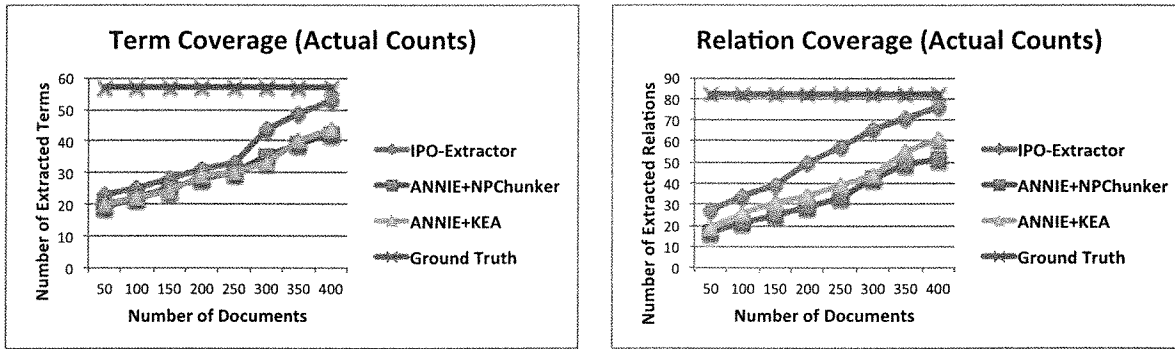


Figure 5.5. Coverage/Correctness Comparison Between Different Methods

The second set of evaluation metrics is used to test the quality of the extraction. We adopt widely accepted evaluation metrics in the Information Retrieval (IR) field, such as precision/recall/F-score. To compute the values of the evaluation metrics, we define the selected terms as *positives* and the dropped term candidates as *negatives*. Since our data is highly skewed – negative sentiment indicators appear much more than positive ones, thus, a precision/recall/F-measure test is appropriate. Due to the fact that traditional precision/recall/F-measure is limited in terms of taking semantics of extractions into account (Meijer et al., 2014), we additionally developed a set of measures to evaluate the quality of term extractions, namely *Semantic Precision* (SP), *Semantic Recall* (SR), and *Semantic F-score* (SF). These metrics are defined as follows (Equation (8) – (10)):

$$SP = \frac{1}{\sqrt{m^2 + n^2} + 1} \left[\sum_{i=1}^m \text{sim}(X, X_i^{\text{Child}}) \times \text{prec}(X_i^{\text{Child}}) + \sum_{j=1}^n \text{sim}(X, X_j^{\text{Parent}}) \times \text{prec}(X_j^{\text{Parent}}) + \text{prec}(X) \right] \quad (8)$$

$$SR = \frac{1}{\sqrt{m^2 + n^2} + 1} \left[\sum_{i=1}^m \text{sim}(X, X_i^{\text{Child}}) \times \text{recall}(X_i^{\text{Child}}) + \sum_{j=1}^n \text{sim}(X, X_j^{\text{Parent}}) \times \text{recall}(X_j^{\text{Parent}}) + \text{recall}(X) \right] \quad (9)$$

$$SF = \frac{2 \times SP \times SR}{SP + SR} \quad (10)$$

In above equations, X_i^{Child} and X_j^{Parent} are the $i^{\text{th}}/j^{\text{th}}$ child/parent of an extracted term X . The children/parents of terms can be retrieved from domain ontology through both taxonomical and non-taxonomical relations (i.e. “part-of” relations indicating taxonomical relations, while “reversedValue” relations indicating non-taxonomical relations; while m and n are the numbers of children/parents of X , respectively. $\text{sim}(X, X_i^{\text{Child}})$ is the semantic similarity (here we use the *WuP* measure) between X and its i^{th} child. $\text{prec}(X)$ and $\text{recall}(X)$ are the traditional precision/recall of X .

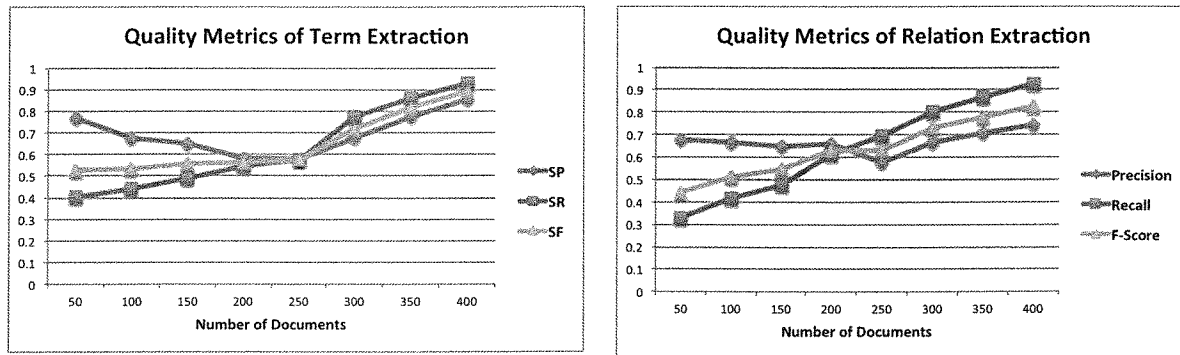


Figure 5.6. Quality Metrics of IPO-Extractor on Select Text Corpus

For evaluating the quality of relation extraction, we directly employ traditional precision/recall/F-score metrics. The SP, SR, and SF of *IPO-Extractor* is shown on the left-side of Figure 5.6 above, while the precision, recall, and F-score of *IPO-Extractor* on different corpus size are shown on the right side of Figure 9 below. With respect to term extraction, *IPO-Extractor* achieves a SP of 85.5%, a SR of 93.0%, and a SF of 89.1%; on the other hand, *IPO-Extractor* achieves a precision of 74.5%, a recall of 92.7%, and an F-score of 82.6% when extracting semantic relations from the select text corpus. Such performances are better than existing mainstream Information Extraction toolkits (*OLLIE*, *REVERB*, *TextRunner*) with respect to term extraction (Gaeta et al., 2011; N. Kang et al., 2011; Ruiz-Martínez, Valencia-García, Martínez-Béjar, & Hoffmann, 2012) and relation extraction (Corro & Gemulla, 2013; Fader et al., 2011).

For the purpose of comparing accuracies of our approach and GATE-native approaches (with help from OLLIE for relation extraction), we have compared the values of recall using different corpus sizes. The comparison results are shown in Figure 5.7 below.

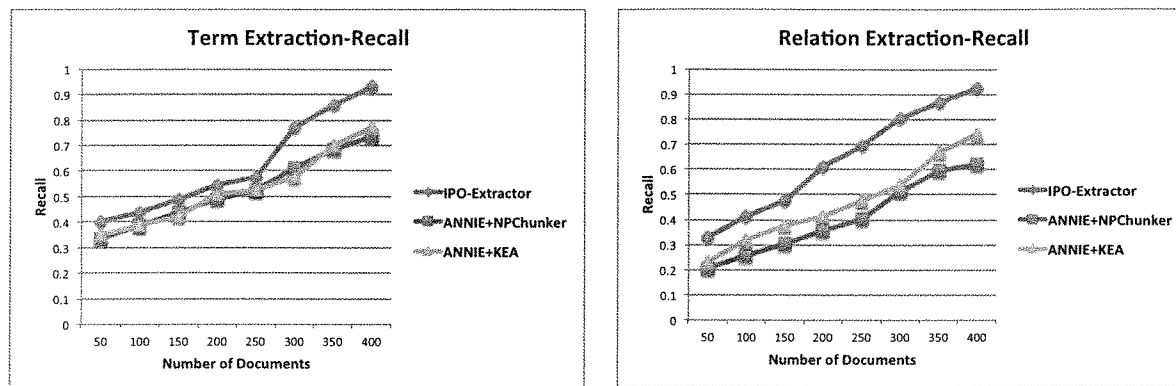


Figure 5.7. Quality Comparison Between Different Approaches

As illustrated in Figure 5.7, *IPO-Extractor* exhibits an evident improvement in terms of extraction quality (measured in recall) with respect to both term and relation extractions, comparing to aforementioned GATE-native extraction approaches. *IPO-Extractor* produces 1.07 – 1.33 times more correct term extractions than *ANNIE+KEA*, the better-performing GATE-based method; while the ratio ranges from 1.25 – 1.5, in terms of relation extractions. Such improvements in recall are due to: a) our proposed approach leverages domain-dependent knowledge; thus, it retrieves more domain-specific terms and relations; b) our proposed approach operates at clause-level, which is finer-grained; such setting helps our relation extraction method yields more accurate results.

However, we have noticed the relatively low precision of *IPO-Extractor* in terms of relation extractions (74.5%). The extraction errors of this kind are because of the erroneous parsing of the dependencies in the clauses (thus affects the identification of the clause types). We will continue to seek more accurate tool(s) for analyzing the grammar trees in clauses, to better understanding the syntactical dependencies within them.

5.2 Predictive Modeling Module

Drawing on the needs of this study and the research gaps identified in Section 2, we articulate the following design requirements for the *FOCAS-IE* (feature-oriented, context-aware,

sentence level Information Extration) analytical pipeline: (1) *Parsing the filings* - The first step of parsing is to read the HTML tags in the documents, the second step is to conduct a deep NLP parsing to identify the linguistic elements (i.e. *tokens/words, sentences*) in the textual contents of the filings. (2) *Segmentation of the documents* – Segmentation involves partition key relevant sections within each prospectus, including the *Risk Factors* and the *MD&A* sections. (3) *Named entity recognition* - Three types of named entities need to be identified in this study, namely, *features, forward-looking indicators, and sentiment-indicative words*. Features refer to the major factors discussed in the *Risk Factors* and the *MD&A* sections that affect the IPO pricing strategy. Forward-looking indicators ensure that the sentences are discussing about the future outlook. Sentiment-indicative words denote the attitudes of the issuer/underwriters toward certain feature(s). (4) *Relationship recognition*: We need to further discover the relationships between the three types of entities mentioned. We are interested in two levels of relationships for this study: the first level relationships are co-occurrence relationships, which imply that at least one instance of each of three types of terms appears at a certain contextual level (e.g., *sentence-level*). The second level relationship is modification, implying that the modifiers, such as the forward-looking indicator(s) and the sentiment-oriented word(s) in the *MD&A* sections are indeed describing a certain feature and they co-occur in a context.

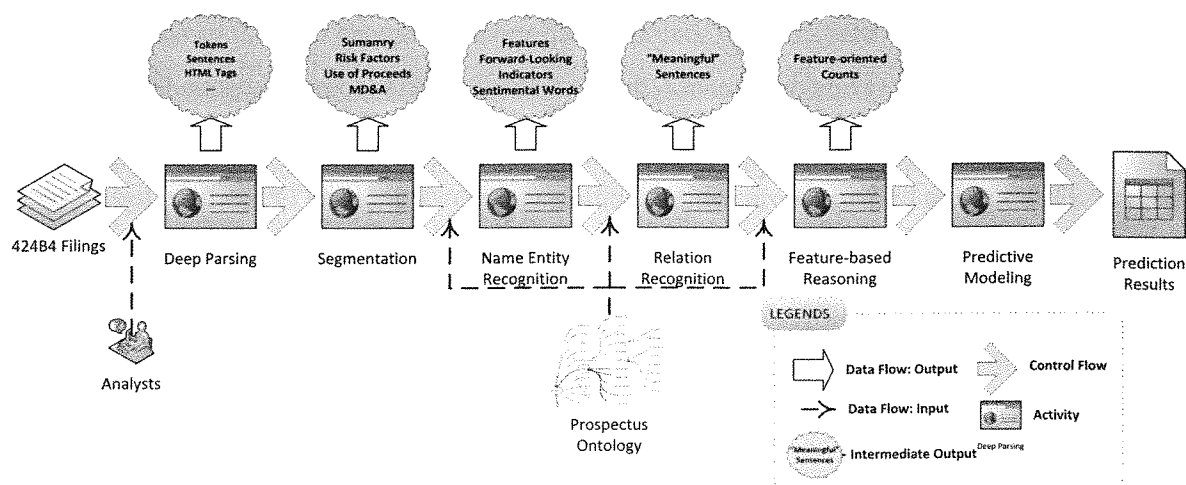


Figure 5.8. FOCAS-IE Analytical Pipeline

The *FOCAS-IE* analytical pipeline is depicted in Figure 5.8. The framework is based on GATE (General Architecture for Text Engineering) as the linguistic platform. *Deep Parsing* is conducted using GATE's native IE system namely *ANNIE*, while *Segmentation* is conducted using GATE's pattern-matching rule engine, namely *JAPE*.

We developed an underlying knowledge structure, namely the '*IPO Ontology*', for recognizing relevant name entities from the annotated sections. On the part of the *MD&A* sections, the IPO Ontology has three major classes, namely features, forward-looking indicators, and sentiment-oriented words. Financial domain experts were asked to help us building the seed concept list for the IPO Ontology, then ontology enrichment techniques are applied for ensuring accuracy and coverage. We have five features in this part of the study: *sales/revenues*, *costs/losses*, *investments*, *net incomes*, and *cash flows*. Related terms of each feature are listed in the prospectus ontology as properties; a total of 50 terms are included for all the 5 features together after the ontology enrichment (e.g., *market shares/earnings belongs in sales/revenues*). 25 terms indicating the prediction/anticipation of future situations are categorized under the class of forward-looking indicators (e.g., *foresee*, *estimate*, *expect*). As for the sentiment-indicative words, we use three word lists from Loughran & McDonald (2011) (*L&M Wordlist*), which are *positive*, *negative*, and *uncertain*, respectively. After applying disambiguation and enrichment techniques, the *positive* word list contains 144 words (e.g., *pleasant*, *ideal*, and *honorable*), the *negative* word list contains 746 words (e.g., *dangerous*, *defective*, and *tragic*), and the *uncertain* word list contains 119 words (i.e. *inexact*, *undecided*, and *intangible*). On the other hand, the IPO Ontology contains ontological (taxonomical and semantic) information regarding the *Risk Factors* sections in the IPO prospectuses in the other half of its structure. Similarly, there are three first level classes, namely (risk-) features, (risk-) mentions, and (risk-) indicators. The detailed structure of this half of the IPO Ontology is discussed in the case study in Section 5.1.2.

Based on the IPO Ontology, relevant named entities are identified using GATE language resources as follows. The *Snowball stemmer* is used to pick up the stem of each word; while *APOLDA* and *onto-gazetteers* are used to provide the semantic annotations based on the terms contained in the IPO Ontology. Next, the context-aware relation recognition using JAPE involves identifying the relations among these terms in the context of sentences. A sentence is annotated as "*meaningful*" if and only if at least one instance of each major class (i.e. in the *MD&A* sections: *features*, *forward-looking indicators*, and *sentiment words*; or in the *Risk Factors* sections: *risk-features*, *risk-mentions*, and *risk-indicators*) appears in it. The modification relations are determined through feature-based reasoning using JAPE, i.e. if only the sentimental words appear in the *n bag-of-words* (*n* adjacent words of the target term, in this study, *n* is set to 5) of the feature, we consider the sentimental words modifies the feature. The

grammar tree structure of each sentence (discussed in Section 4.3.5) is also used to detect the modification relationships. We ensure the quality of the *FOCAS-IE* analytical pipeline through a hybrid (automatic and manual) checking approach, incorporating exception handling where needed.

The *feature-based reasoning* module returns the counts of “*meaningful*” sentences by features and sentiments. The identification of “*meaningful*” sentences in the *Risk Factors* sections and the *MD&A* sections are similar, but different. In the *MD&A* sections, for each 424B4 filing for an IPO, a total of 15 counts are returned (5 features x 3 sentiments). It may be noted that a sentence might overlap over different features/sentiments – for instance, a sentence might discuss both *sales* and *net income*, or contain a transition structure where the first clause is denoting a positive sentiment while the second one is negative. The counts are then normalized to account for variation in lengths of different *MD&A* sections. In the *Risk Factors* sections, a total of 5 counts (each one corresponding to a unique risk-feature) are returned for each Form 424B4 filing. Note that sentences might overlap over different features – similar to the *MD&A* sections. Following this, the *predictive modeling* module analyzes the directions of the changes (Positive trends, non-Positive trends) in both pre-IPO price adjustments and post-IPO initial returns using different set of predictors (i.e. 5 predictors from the *Risk Factors* sections, 15 predictors from the *MD&A* sections, a combination of both, or aggregated predictors). The predictive modeling is detailed in Section 4.4. In the next section, the dataset used in the experiments need to be discussed in detail.

5.3 Study Data

The research study data mainly includes IPO prospectuses retrieved from the US SEC’s EDGAR system, selected for the companies that went public in the most recent decade (January 1, 2003 – December 31, 2013). In accordance with similar prior studies, we have excluded the following: Financial firms (i.e. banks, loan firms, and insurance companies), American Depositary Receipts (*ADRs*), real estate investment trusts (*REITs*), close-end funds, and firms with offering price less than five dollars (Hanley & Hoberg, 2010; Loughran & McDonald, 2013). We have further restricted our sample by the stock type of *common stock* or *ordinary*

stock. The Center for Research in Security Prices (CRSP) serves as the source for data such as stock offering prices, initial returns, and value-weighted returns of each IPO.

Form 424 is the final and approved version of the IPO prospectuses, which is filed on or within several days after the IPO day. Among all 424B variants, we have selected Form 424B4 in this study. All 424B4 entering the sample must have a valid SEC Central Index key (*CIK*), CRSP *permno*, and key sections including the *Risk Factors* section and the MD&A section. Using an intelligent web crawling algorithm, we retrieved 424B4 filings from SEC EDGAR for 713 US IPOs, which were then filtered through a filtering algorithm to meet aforementioned sampling boundary restrictions, resulting in 513 filings. We believe such sample is comparable to the sample used in similar studies – for instance, Loughran & McDonald collected approximately 453 424B variants from their initial sample (2013). The lengths of the prospectuses and the sections within them indicate computational complexity of the study, the average length of the prospectus is 57,854 words, with an average length of the risk factors sections as 1,812 words, and the MD&A sections as 1,195 words.

5.3.1 Descriptive Statistics

The descriptive statistics for the target variables and predictors are shown in Table 5.2 below.

Table 5.2. Summary Statistics of All Variables

Variable	Mean	Std.dev.	1 st QTR	Median	3 rd QTR	t-statistics
(A) Target Variables						
<i>PRCREV</i>	0.3887	0.4879	0	0	1	18.12
<i>X1stDay</i>	0.1992	0.3998	0	0	0	11.33
(B) Predictors from Prior Studies						
<i>Average Annual Sales</i>	106.39	27.75	18.18	66.02	326.06	31.93
<i>Up Revision</i>	5.62%	10.80%	0	0	7.70%	11.83
<i>Top-Tier dummy</i>	0.7988	0.4013	0	1	1	45.27
<i>Positive EPS dummy</i>	0.7311	0.4438	0	1	1	37.46
<i>Prior Nasdaq 15-day return</i>	0.0041	0.0437	-0.0183	0.0010	0.0033	2.12
<i>Share overhang</i>	4.3840	3.8624	2.9010	3.7690	4.9085	25.81
<i>Days between S-1 and 424</i>	204.43	14.26	85.00	107.00	170.00	45.53
(C) MD&A Predictors						
<i>MDAWeight</i>	9.85%	2.81%	5.67%	9.56%	14.68%	79.60
<i>Sale*Positive</i>	0.1875	0.1039	0.1111	0.1757	0.2549	41.04
<i>Sale*Negative</i>	0.1273	0.0729	0.0714	0.1205	0.1724	39.69
<i>Sale*Uncertain</i>	0.1667	0.0871	0.1045	0.1579	0.2222	43.51

<i>Cost*Positive</i>	0.2831	0.0906	0.2222	0.2769	0.3421	71.08
<i>Cost*Negative</i>	0.2465	0.0919	0.1818	0.2364	0.3030	60.98
<i>Cost*Uncertain</i>	0.2299	0.0850	0.1667	0.2250	0.2821	61.47
<i>Investment*Positive</i>	0.1330	0.0784	0.0787	0.1228	0.1795	38.58
<i>Investment*Negative</i>	0.0937	0.0597	0.0500	0.0847	0.1250	35.67
<i>Investment*Uncertain</i>	0.1459	0.0803	0.0893	0.1356	0.1875	41.31
<i>Net Income*Positive</i>	0.0802	0.0574	0.0435	0.0723	0.1071	31.73
<i>Net Income*Negative</i>	0.1194	0.0714	0.0682	0.1098	0.1616	38.04
<i>Net Income*Uncertain</i>	0.0890	0.0572	0.0488	0.0800	0.1212	35.41
<i>Cash flow*Positive</i>	0.1271	0.0652	0.0825	0.1200	0.1628	44.30
<i>Cash flow*Negative</i>	0.0782	0.0547	0.0395	0.0682	0.1081	32.51
<i>Cash flow*Uncertain</i>	0.1341	0.0683	0.0833	0.1250	0.1739	44.63
(D) Risk Factors Predictors						
<i>RiskWeight</i>	25.09%	47.94%	18.06%	20.93%	24.18%	11.90
<i>Competitiveness</i>	0.0848	0.0378	0.0581	0.0789	0.1077	50.96
<i>Customer</i>	0.2404	0.0936	0.1667	0.2373	0.3085	58.41
<i>Growth</i>	0.7881	0.0894	0.7527	0.8000	0.8442	200.53
<i>Management Skills</i>	0.2166	0.0833	0.1587	0.2093	0.2667	59.08
<i>Lawsuits</i>	0.1384	0.0564	0.1029	0.1351	0.1739	55.83
<i>Stock Prices</i>	0.1718	0.0583	0.1356	0.1613	0.1974	67.03
(E) Aggregated Predictors						
<i>MDAScore</i>	0.0107	0.0101	0.0045	0.0111	0.0171	24.28
<i>RiskScore</i>	-4.1536	0.3757	-4.3284	-4.1782	-4.0345	-251.36

The sample contains 517 424B4 filings with an offer price greater than five dollars per share; each filing represents a successful U.S. IPO with the period between January, 1, 2003 – December, 31, 2013. More information regarding the constructed of the sample and the description of each variable can be found in Table 4.3 above. *MDAWeight* is a variable reflecting the ratio of “*meaningful*” sentences to the total sentences in the MD&A section for each filing. The naming of the predictor variables is explained as feature*sentiment; i.e. *Sale*Positive* corresponds to the counts of “*meaningful*” sentences which are discussing the sales/revenues feature and are of positive sentiment. All the predictor variables are normalized against the count of total “*meaningful*” sentences in the MD&A section of the same filing. Similar definitions apply to the *risk-features*. The average annual sales of the firms are in the unit in millions of dollars.

Summary statistics in Table 5.2 illustrate several informative facts. Panel (A) presents the descriptive statistics of the target variables. It is clear that compared to the post-IPO price changes, the pre-IPO price changes are closer to the positive end – meaning issuers/underwriters

tend to revise the offering price upward. This is consistent with the “underpricing” phenomena observed and reported in the finance literature (i.e. (Ljungqvist et al., 2006; Lowry, 2003)).

Table 5.2 also reports the facts regarding the IPO characteristics in our sample (in Panel (B)). Firstly, consistent with prior related studies, the issuing firms are relatively small, with an average trailing annual sale of \$106.39 million. The average upward revision from the mid point of the pricing range to the final offering price is 5.62%, which is significantly lower than the sample selected in (Loughran & McDonald, 2013) – it is because we exclude the Internet Bubble era in our sample. Also because of the same exclusion of the data, our sample has a positive trailing EPS values. On average, 79.88% of the companies in the sample (413 companies) were taken public by a top-tier underwriter. Large variation exists in the days between the filing date of the initial S-1 and the final 424B4. On average, 204.41 days separate the two dates, while the 1st Quarter is 85 days and the 3rd Quarter is 170 days.

Panel (C) of Table 5.2 reports the summary statistics of the predictors from the *MD&A* sections. We first calculate the percentage of “*meaningful*” sentences with respect to the overall total sentences in the *MD&A* section (*MDAWeight*). On average, 9.85% (standard deviation: 2.81%) of the total sentences in the *MD&A* sections are categorized as “*meaningful*” (feature-oriented, forward-looking, and sentiment-indicative), with a similar median of 9.56%. We further break the “*meaningful*” sentences down by the selected features (*sales/revenues*, *costs/losses*, *investments*, *net income*, and *cash flows*) and three sentiments (*positive*, *negative*, and *uncertain*). 15 variables representing all the pairs of (*feature*, *sentiment*) are computed (e.g., the percentage of forward-looking sentences discussing the feature *sales/revenues* with the *positive* sentiment is denoted as *Sale*Positive*). The descriptive statistics indicate that the percentages of the feature *costs/losses* are relatively higher than other features, which suggests that managers tend to discuss more about the costs and losses of the company’s future operations. Also, four out of five features have more *positive* sentiments than *negative* sentiments (except for *net income*), which is consistent with the findings in Hanley and Hoberg (2010) that the *MD&A* sections are generally positive in tone. The exception of ‘net income’ feature suggests that managers tend to be more conservative discussing future incomes.

Panel (D) of Table 5.2 reports the summary statistics of the predictors from the *Risk Factors* sections. It is clear that the issuers/underwriters have substantial concerns regarding the risk factors related to the *growth* of the organization (an average weight of 78.81%). They also

highlight risk factors regarding the *customers* (24.04%) and *management skills* (21.66%). This is consistent with the common understanding and the financial literature – these three are very important factors impacting the health of the business. Similarly, the variable *RiskWeight* indicates the relative length of the *Risk Factors* sections – on average they take up 25.09% of the overall length of the prospectus. Besides the regulatory requirements from SEC, this observation is also consistent with the conservatism of the underwriters.

The last panel of Table 5.2, Panel (E), presents the descriptive statistics of the aggregated predictors, namely *RiskScore* and *MDAScore*. One thing worth noting is that *RiskScore* has a negative mean – meaning it has a negative impact on IPO pricing; while the *MDAScore* has a positive mean; these observations are also consistent with prior related studies – both scores are calculated via Equation (6) and (7) in Chapter 4.

5.4 Experiments, Results and Discussions

In this section, firstly the designs of the experiments are elaborated. The experiments are designed to predict both pre-IPO price revisions and post-IPO price changes using the selected sample discussed in Section 5.3. Four distinguished experiment alternatives are presented in this study in order to test the predictive power of different sets of predictors, at variant abstraction levels. The four alternatives are as follows:

- Alternative 1: This alternative examines whether the *risk-features* possess predictive powers toward pre-IPO price revisions and post-IPO price changes. Predictive modeling techniques mentioned in Section 4.4.3 are used in this alternative. Predictors from prior related studies are included in each model as well. The training data set is used to train the models; the validation data set is used to compare across the models; the testing data set is used to final assess the selected model from prior steps.
- Alternative 2: This alternative examines whether the *MD&A-features* possess predictive powers toward pre-IPO price revisions and post-IPO price changes. Predictive modeling techniques mentioned in Section 4.4.3 are used in this alternative. Predictors from prior related studies are included in each model as well. The training data set is used to train the models; the validation data set is used to compare across the models; the testing data set is used to final assess the selected model from prior steps.

- Alternative 3: This alternative examines whether both the *risk-features* and the *MD&A-features* possess predictive powers toward pre-IPO price revisions and post-IPO price changes. Predictive modeling techniques mentioned in Section 4.4.3 are used in this alternative. Predictors from prior related studies are included in each model as well. The training data set is used to train the models; the validation data set is used to compare across the models; the testing data set is used to final assess the selected model from prior steps.
- Alternative 4: This alternative examines whether the aggregated *risk-features* and *MD&A-features*, namely *RiskScore* and *MDAScore*, possess predictive powers toward pre-IPO price revisions and post-IPO price changes. This alternative is designed for dimension reduction purposes. Predictive modeling techniques mentioned in Section 4.4.3 are used in this alternative. Predictors from prior related studies are included in each model as well. The training data set is used to train the models; the validation data set is used to compare across the models; the testing data set is used to final assess the selected model from prior steps.

Details toward the four alternatives are discussed in Section 5.4.1 through 5.4.4.

Predictive Modeling techniques detailed in Section 4.4.2 (listed in Table 4.4) are used in each experiment alternative.

Results validation and communication are other two main tasks in this section. Firstly, the results from aforementioned experiments are demonstrated; then they are validated along two dimensions: accuracy and performance. Well-accepted data analytics evaluation metrics and techniques (discussed in detail in Section 4.4.3) are adopted for the results. Then the results are interpreted, with respect to the hypotheses associated with each experiment, and with assured relevance of the finance domain and the IPO process. The result validations are discussed along with each alternative; while the result communication is discussed in Section 5.4.5.

5.4.1. Alternative 1: Using Risk Features and Predictors from Prior Studies

In this alternative, the predictive power of *risk-features*, in combination with the predictors from prior studies, toward pre- and post-IPO pricing trends, is examined.

Firstly, pre-IPO price revision (*PRCREV*) is selected as the target variable in the predictive model. The *risk-feature* predictors are also selected, along with the predictors from prior studies – excluding *Up Revision*. The rationale behind this design decision is *Up Revision* is highly correlated with the target variable (*PRCREV*), which would compromise the prediction.

As discussed in Section 4.4.3, oversampling is used for the imbalanced *PRCREV* in the training data set. Thus, the training data set has 436 data points, the validation data set has 92 data points, and the testing data set has 84 data points.

Table 5.3(a) shows the confusion matrices, which provide evidences for computing the accuracy metrics (precision, recall, accuracy, F-score) as discussed in Section 4.4.3.

Table 5.3(a). Confusion Matrices of PRCREV in Alternative 1

PRCREV	Training Data Set				Validation Data Set			
Model	FP	TP	FN	TN	FP	TP	FN	TN
DT	99	180	24	133	7	34	27	24
RF	29	122	82	203	7	34	27	24
EN	36	150	54	196	11	36	25	20
ANN	40	143	61	192	12	37	24	19
LR	63	113	91	169	11	34	27	20
GLM	63	113	91	169	11	34	27	20
SVM	44	95	109	188	9	31	30	21
k-NN	35	121	83	197	12	30	31	19

In Table 5.3(a), FP refers to False Positive (incorrectly classified to *PRCREV* = 1), TP refers to True Positive (correctly classified to *PRCREV* = 1); while FN refers to False Negative (incorrectly classified to *PRCREV* = 0), TN refers to True Negative (correctly classified to *PRCREV* = 0). Based on Table 5.3(a) and Equations 8a – 8d, the accuracy metrics of each predictive models using *PRCREV* in Alternative 1 is computed and illustrated in Table 5.3(b).

Table 5.3(b). Accuracy Metrics of PRCREV in Alternative 1

PRCREV	Training Data Set				Validation Data Set			
Model	Precision	Recall	Accuracy	F-Score	Precision	Recall	Accuracy	F-Score
DT	0.645	0.882	0.718	0.745	0.829	0.557	0.630	0.667
RF	0.808	0.598	0.745	0.687	0.829	0.557	0.630	0.667
EN	0.806	0.735	0.794	0.769	0.766	0.590	0.609	0.667
ANN	0.781	0.701	0.768	0.739	0.755	0.607	0.609	0.673
LR	0.642	0.554	0.647	0.595	0.756	0.557	0.587	0.642
GLM	0.642	0.554	0.647	0.595	0.756	0.557	0.587	0.642
SVM	0.683	0.466	0.649	0.554	0.775	0.508	0.571	0.614
k-NN	0.776	0.593	0.729	0.672	0.714	0.492	0.533	0.583

The accuracy metrics of the predictive models with the training data set illustrate how the model is trained (shown on the left four columns of Table 5.3(b)); while the accuracy metrics of the predictive models with the validation data set (shown on the right four columns of Table 5.3(b)) demonstrate the accuracies of the trained models. As discussed in Section 4.4.3, F-score

of the validation data set (most right column in Table 5.3(b)) is determined as the metric for selecting the predictive model for final assessment: thus, ANN is selected as the most accurate model of Alternative 1 with *PRCREV* as the target variable.

On the other hand, the efficiency metrics (Lift and AUC) are also examined for the predictive models, with training and validation data sets. The results are reported in Table 5.3(c) below.

Table 5.3(c) Efficiency Metrics of PRCREV in Alternative 1

PRCREV	Training Data Set		Validation Data Set	
Model	Lift	AUC	Lift	AUC
DT	1.694	0.764	1.241	0.587
RF	1.666	0.792	1.213	0.631
EN	1.751	0.843	1.820	0.660
ANN	1.794	0.832	1.517	0.640
LR	1.495	0.721	1.213	0.612
GLM	1.495	0.721	1.213	0.612
SVM	1.495	0.710	0.910	0.630
k-NN	1.820	0.813	1.181	0.549

The AUC diagram of the predictive models with the training, validation, and testing data sets are shown in Figure 5.9 below (from left to right).

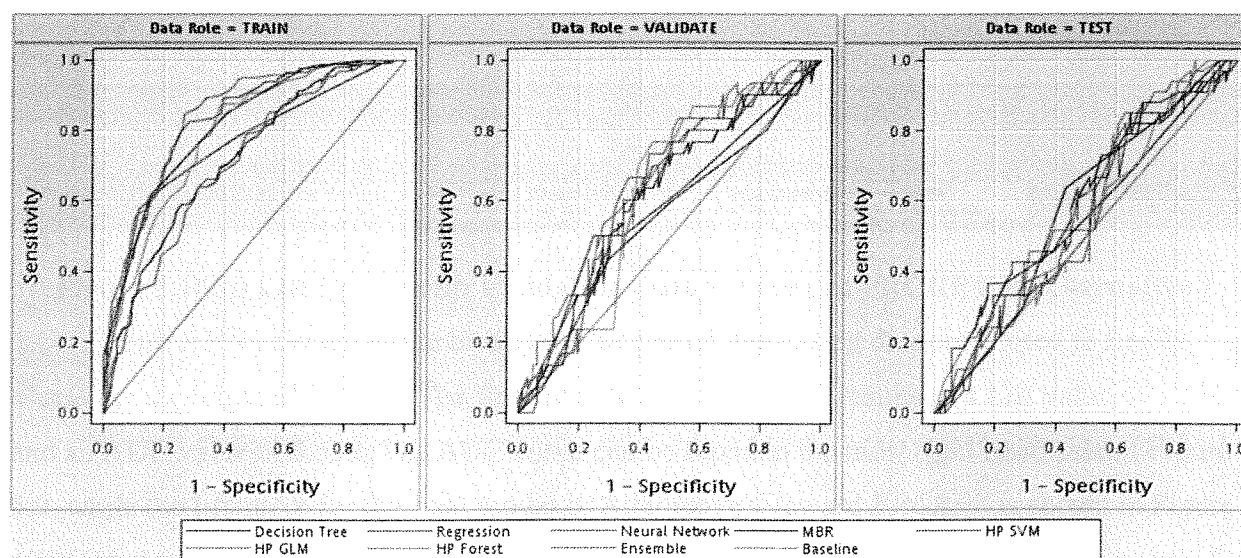


Figure 5.9. AUC Diagram of Predictive Models toward PRCREV in Alternative 1

The Lift Curves for predictive models in Alternative 1 toward *PRCREV* are shown in Figure 5.10 (a) and (b) below.

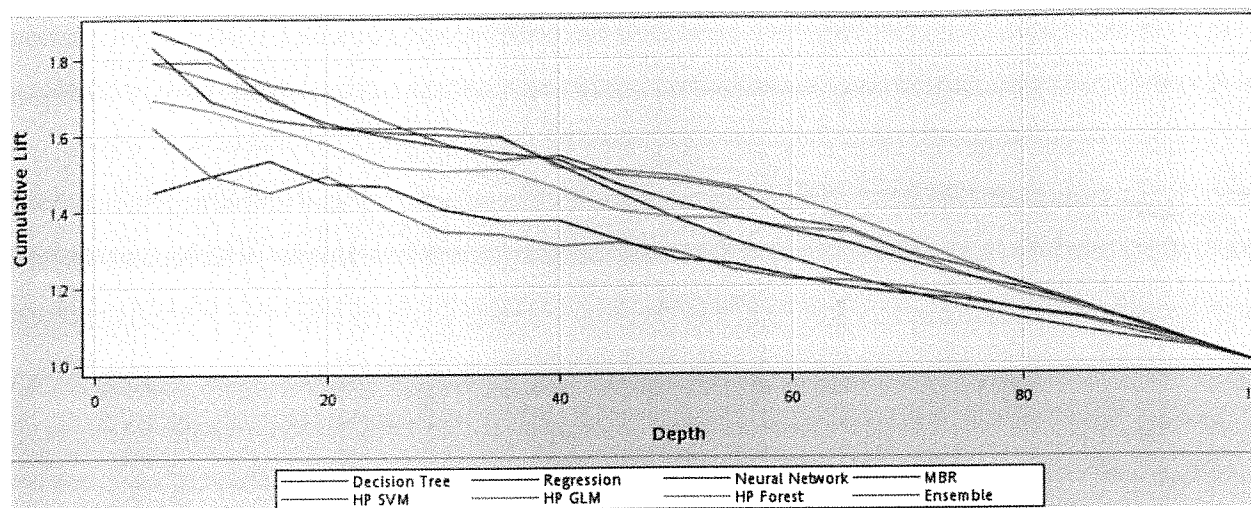


Figure 5.10(a) Lift Curve with Training Data Set toward PRCREV in Alternative 1

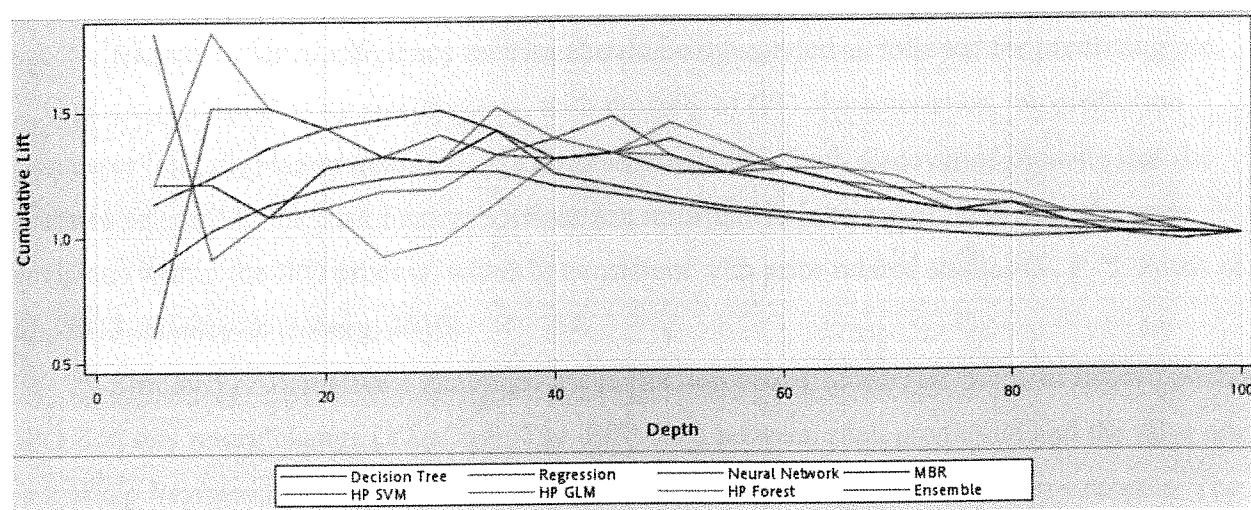


Figure 5.10(b) Lift Curve with Validation Data Set toward PRCREV in Alternative 1

With ANN selected as the most accurate model, the testing date set is used to finally assess the modeling efficiency. The final selected model (ANN) results in a Lift value of 1.131 and an AUC value of 0.559. The AUC curve can be found in the most right hand side of Figure 5.9, while the lift curve for the final model in Alternative 1 toward *PRCREV* is shown in Figure 5.10(c) below.

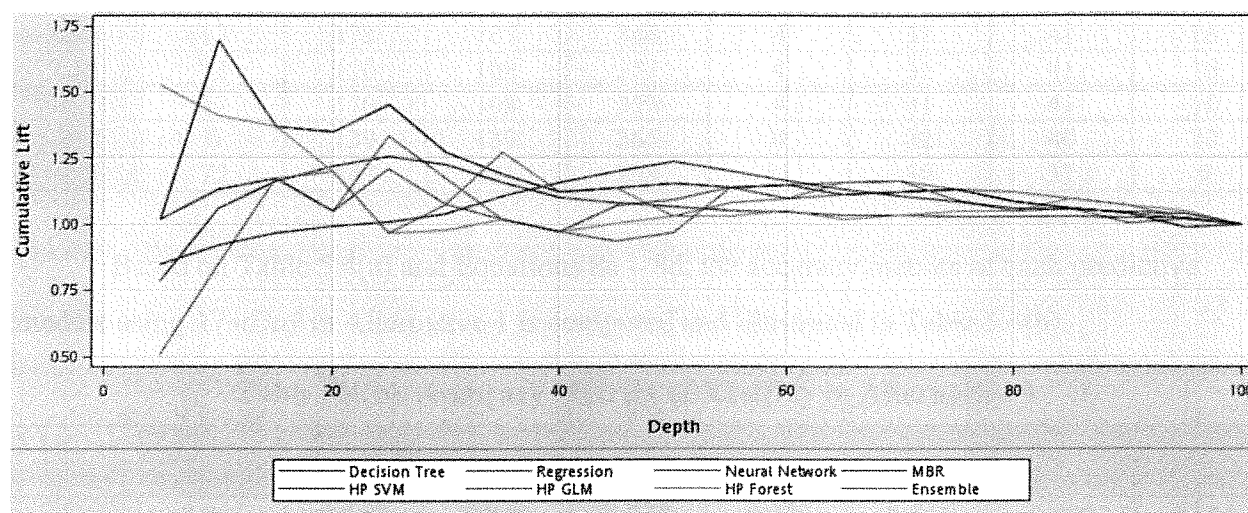


Figure 5.10(c) Lift Curve with Testing Data Set toward PRCREV in Alternative 1

Aforementioned efficiency metrics should be interpreted as follows: for predicting pre-IPO pricing revision using *risk-features*, with the help of IPO characteristics, the prediction possesses (slightly) better quality than a random prediction. Such discoveries indicate that the features selected in the Risk Factors sections are informative in terms of predicting the price revisions within the IPO process, which is consistent with prior related studies (S. P. S. Ferris et al., 2013; Hanley & Hoberg, 2010).

Similarly, Alternative 1 examines the prediction power of the *risk-features* toward post-IPO first day return change (*X1stDay*). The difference between such prediction and the prior one is that *Up Revision*, as a significant IPO characteristic, is included in the predictive models. The reason is that the high correlation no longer exists between *Up Revision* and *X1stDay*. Similar oversampling treatment is conducted to the training data set, results in a training data set with 643 data points, a validation data set with 92 data points, and a testing data set with 84 data points.

Table 5.4(a) shows the confusion matrices, which provide evidences for computing the accuracy metrics (precision, recall, accuracy, F-score) as discussed in Section 4.4.3.

Table 5.4(a). Confusion Matrices of X1stDay in Alternative 1

X1stDay	Training Data Set				Validation Data Set			
Model	FP	TP	FN	TN	FP	TP	FN	TN
DT	5	180	98	360	3	39	33	17
RF	9	129	149	356	7	22	50	13
EN	0	176	102	365	2	39	33	17

ANN	17	171	107	348	4	31	41	16
LR	12	170	108	353	4	31	41	16
GLM	12	170	108	353	4	31	41	16
SVM	0	139	139	365	0	32	40	20
k-NN	72	147	131	293	11	35	37	9

Based on Table 5.4(a) and Equations 8a – 8d, the accuracy metrics of each predictive models using *X1stDay* in Alternative 1 is computed and illustrated in Table 5.4(b).

Table 5.4(b). Accuracy Metrics of X1stDay in Alternative 1

X1stDay	Training Data Set				Validation Data Set			
Model	Precision	Recall	Accuracy	F-Score	Precision	Recall	Accuracy	F-Score
DT	0.973	0.647	0.840	0.778	0.929	0.542	0.609	0.684
RF	0.935	0.464	0.754	0.620	0.759	0.306	0.380	0.436
EN	1.000	0.633	0.841	0.775	0.951	0.542	0.615	0.690
ANN	0.910	0.615	0.807	0.734	0.886	0.431	0.511	0.579
LR	0.934	0.612	0.813	0.739	0.886	0.431	0.511	0.579
GLM	0.934	0.612	0.813	0.739	0.886	0.431	0.511	0.579
SVM	1.000	0.500	0.784	0.667	1.000	0.444	0.565	0.615
k-NN	0.671	0.529	0.684	0.592	0.761	0.486	0.478	0.593

The accuracy metrics of the predictive models with the training data set illustrate how the model is trained (shown on the left four columns of Table 5.4(b)); while the accuracy metrics of the predictive models with the validation data set (shown on the right four columns of Table 5.4(b)) demonstrate the accuracies of the trained models. As discussed in Section 4.4.3, F-score of the validation data set (most right column in Table 5.4(b)) is determined as the metric for selecting the predictive model for final assessment: thus, EN is selected as the most accurate model of Alternative 1 with *X1stDay* as the target variable. It is worth noting that the prediction toward *X1stDay* is comparable toward *PRCREV*, indicating that a) *risk-features* are equivalently significant toward both pre- and post-IPO pricing trends; and b) pre-IPO price revisions (operationalized in *Up Revision*), has a substantial effect on post-IPO pricing trends.

On the other hand, the efficiency metrics (Lift and AUC) are also examined for the predictive models, with training and validation data sets. The results are reported in Table 5.4(c) below. It also proves the selection of EN as the final model.

Table 5.4(c) Efficiency Metrics of X1stDay in Alternative 1

X1stDay	Training Data Set		Validation Data Set	
Model	Lift	AUC	Lift	AUC
DT	1.400	0.830	1.402	0.677

RF	1.650	0.870	0.958	0.488
EN	1.630	0.890	0.958	0.702
ANN	1.710	0.840	1.437	0.681
LR	1.600	0.850	1.437	0.626
GLM	1.600	0.850	1.437	0.626
SVM	1.270	0.810	0.479	0.666
k-NN	1.470	0.760	0.000	0.439

The AUC diagram of the predictive models with the training, validation, and testing data sets are shown in Figure 5.11 below (from left to right).

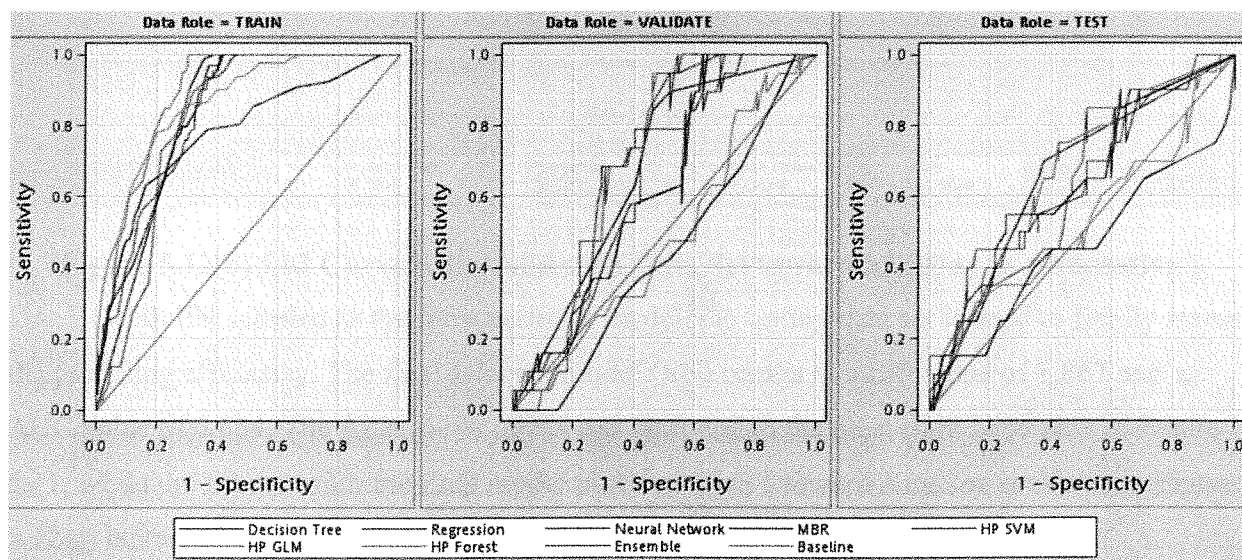


Figure 5.11. AUC Diagram of Predictive Models toward X1stDay in Alternative 1

The Lift Curves for predictive models in Alternative 1 toward *X1stDay* are shown in Figure 5.12 (a) and (b) below.

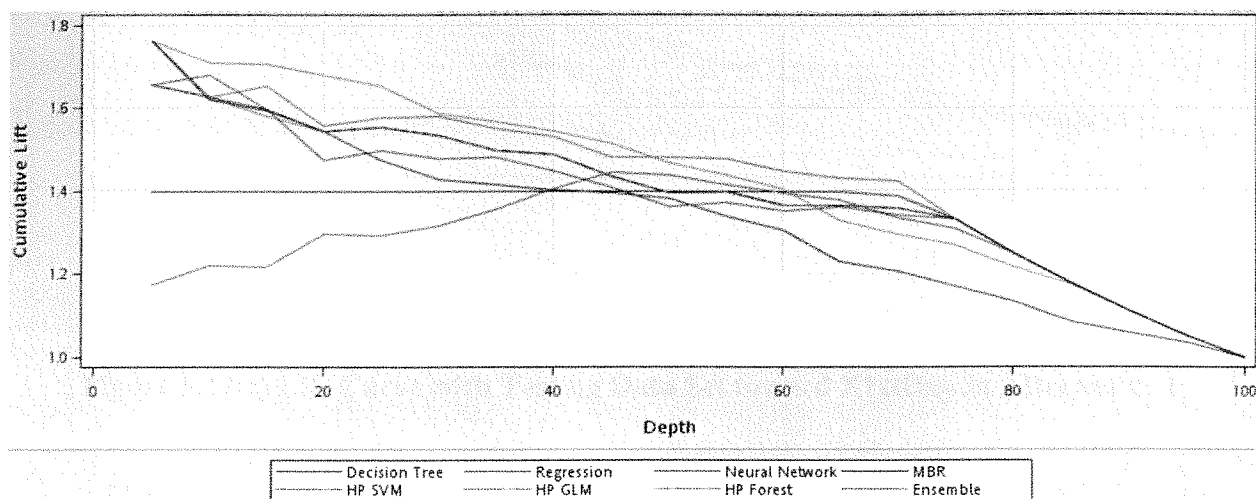


Figure 5.12(a) Lift Curve with Training Data Set toward X1stDay in Alternative 1

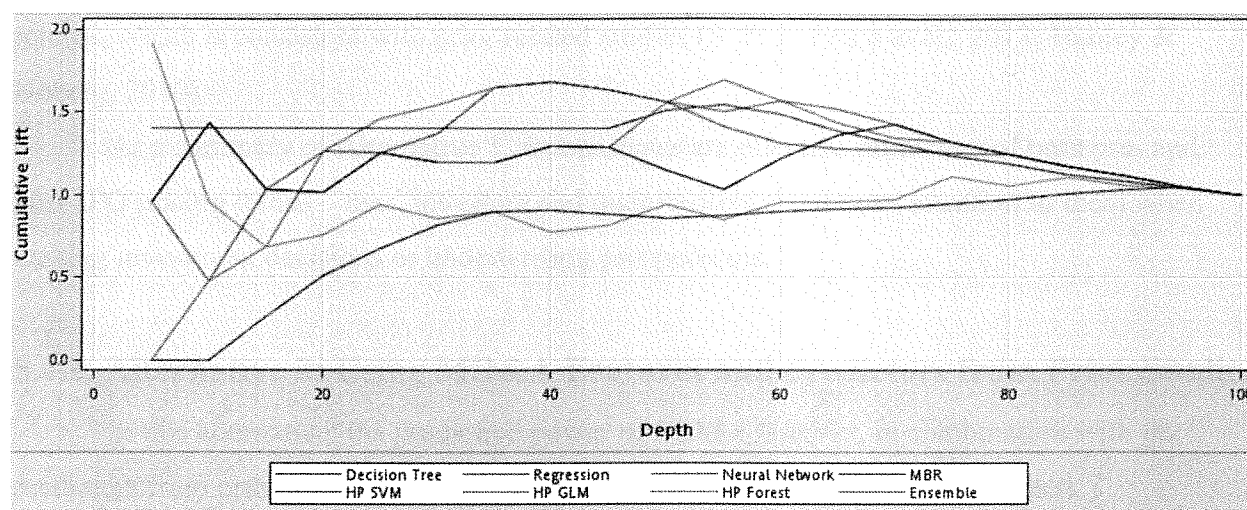


Figure 5.12(b) Lift Curve with Validation Data Set toward X1stDay in Alternative 1

With EN selected as the most accurate model, the testing date set is used to finally assess the modeling efficiency. The final selected model (EN) results in a Lift value of 1.867 and an AUC value of 0.664. The AUC curve of EN can be found in the most right hand side of Figure 5.11, while the lift curve for the final model in Alternative 1 toward *X1stDay* is shown in Figure 5.12(c) below.

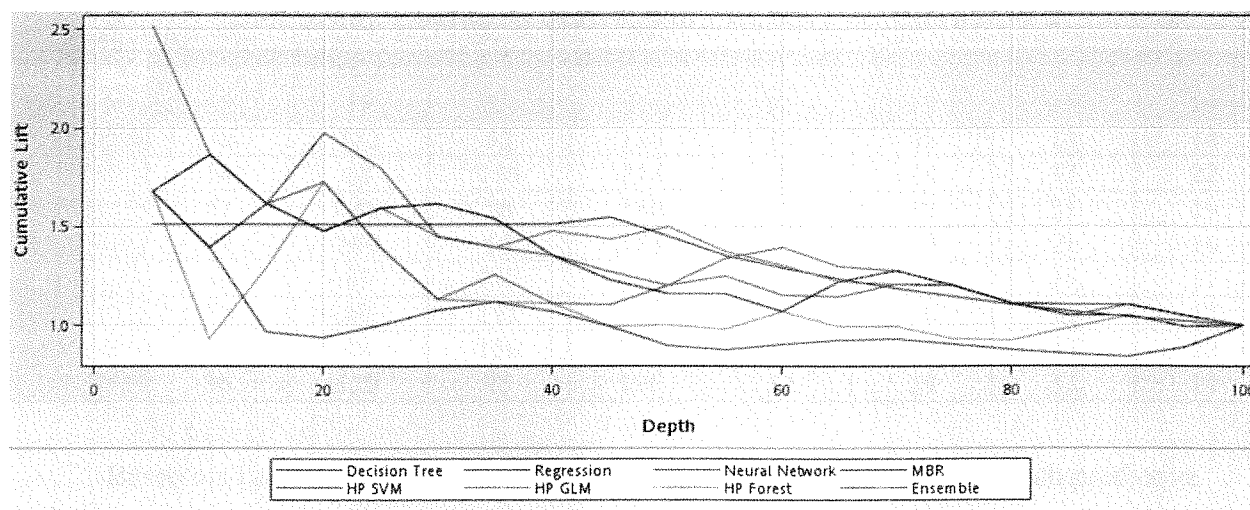


Figure 5.12(c) Lift Curve with Testing Data Set toward X1stDay in Alternative 1

Aforementioned efficiency metrics should be interpreted as follows: for predicting post-IPO pricing revision using *risk-features*, with the help of IPO characteristics, the prediction

possesses better quality than a random prediction. Such discoveries indicate that the features selected in the *Risk Factors* sections are informative in terms of predicting the first day initial returns, which is consistent with prior related studies (S. P. S. Ferris et al., 2013; Hanley & Hoberg, 2010).

As a summary of Alternative 1, *risk-features* are effective predictors of both pre- and post-IPO pricing trends – thus, investors and underwriters should pay attention to them when making investment decisions or underwriting prospectuses.

5.4.2. Alternative 2: Using MD&A Features and Predictors from Prior Studies

In this alternative, the predictive power of *MD&A features*, in combination with the predictors from prior studies, toward pre- and post-IPO pricing trends, is examined.

Firstly, pre-IPO price revision (*PRCREV*) is selected as the target variable in the predictive model. The *MD&A feature* predictors are also selected, along with the predictors from prior studies – excluding *Up Revision*. Similar oversampling treatment is conducted, as described in Alternative 1, yielding same resulting data points in the partitioned data sets.

Table 5.5(a) shows the confusion matrices, which provide evidences for computing the accuracy metrics (precision, recall, accuracy, F-score) as discussed in Section 4.4.3.

Table 5.5(a). Confusion Matrices of PRCREV in Alternative 2

PRCREV	Training Data Set				Validation Data Set			
Model	FP	TP	FN	TN	FP	TP	FN	TN
DT	106	172	32	126	17	48	13	14
RF	14	68	136	218	9	24	37	22
EN	52	148	56	180	12	40	21	19
ANN	49	150	54	183	15	40	21	16
LR	50	89	115	182	14	26	35	17
GLM	50	89	115	182	14	26	35	17
SVM	22	48	156	210	8	18	43	22
k-NN	36	129	75	196	12	29	32	19

Based on Table 5.5(a) and Equations 8a – 8d, the accuracy metrics of each predictive models using *PRCREV* in Alternative 2 is computed and illustrated in Table 5.5(b).

Table 5.5(b). Accuracy Metrics of PRCREV in Alternative 2

PRCREV	Training Data Set				Validation Data Set			
Model	Precision	Recall	Accuracy	F-Score	Precision	Recall	Accuracy	F-Score
DT	0.619	0.843	0.683	0.714	0.738	0.787	0.674	0.762

RF	0.829	0.333	0.656	0.476	0.727	0.393	0.500	0.511
EN	0.740	0.725	0.752	0.733	0.769	0.656	0.641	0.708
ANN	0.754	0.735	0.764	0.744	0.727	0.656	0.609	0.690
LR	0.640	0.436	0.622	0.519	0.650	0.426	0.467	0.515
GLM	0.640	0.436	0.622	0.519	0.650	0.426	0.467	0.515
SVM	0.686	0.235	0.592	0.350	0.692	0.295	0.440	0.414
k-NN	0.782	0.632	0.745	0.699	0.707	0.475	0.522	0.569

The accuracy metrics of the predictive models with the training data set illustrate how the model is trained (shown on the left four columns of Table 5.5(b)); while the accuracy metrics of the predictive models with the validation data set (shown on the right four columns of Table 5.5(b)) demonstrate the accuracies of the trained models. As discussed in Section 4.4.3, F-score of the validation data set (most right column in Table 5.5(b)) is determined as the metric for selecting the predictive model for final assessment: thus, DT is selected as the most accurate model of Alternative 2 with *PRCREV* as the target variable.

On the other hand, the efficiency metrics (Lift and AUC) are also examined for the predictive models, with training and validation data sets. The results are reported in Table 5.5(c) below.

Table 5.5(c) Efficiency Metrics of PRCREV in Alternative 2

PRCREV	Training Data Set		Validation Data Set	
Model	Lift	AUC	Lift	AUC
DT	1.579	0.754	2.085	0.645
RF	1.708	0.836	2.123	0.637
EN	1.708	0.836	2.123	0.633
ANN	1.845	0.826	1.141	0.555
LR	1.708	0.759	1.921	0.614
GLM	1.068	0.635	1.517	0.532
SVM	1.068	0.635	1.517	0.532
k-NN	1.068	0.655	1.213	0.517

The AUC diagram of the predictive models with the training, validation, and testing data sets are shown in Figure 5.13 below (from left to right).

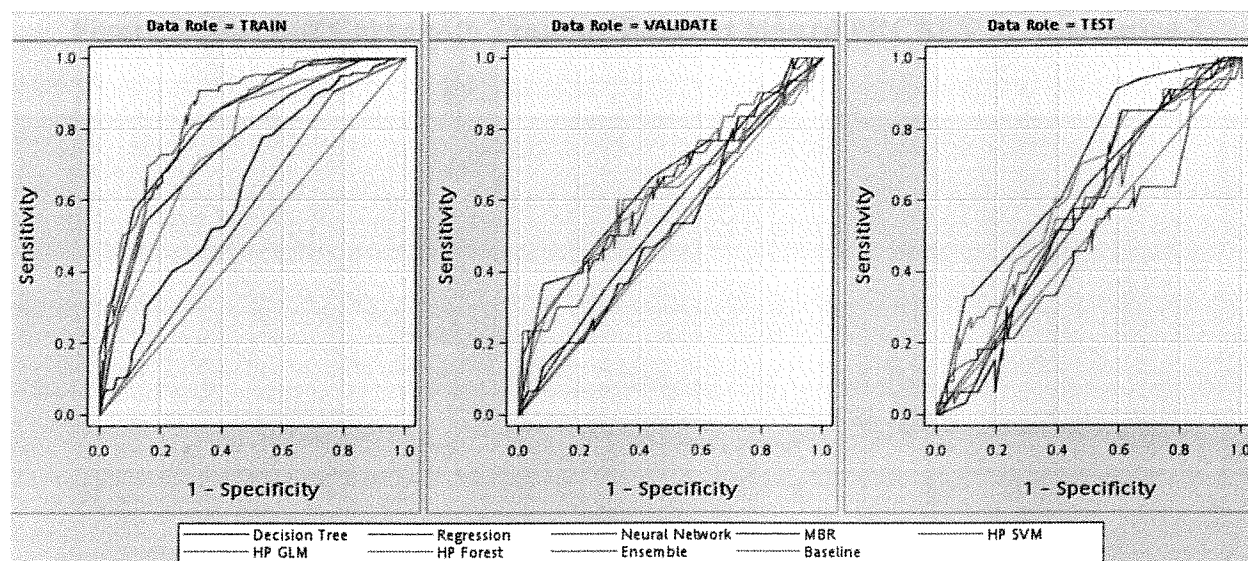


Figure 5.13. AUC Diagram of Predictive Models toward PRCREV in Alternative 2

The Lift Curves for predictive models in Alternative 2 toward *PRCREV* are shown in Figure 5.14 (a) and (b) below.

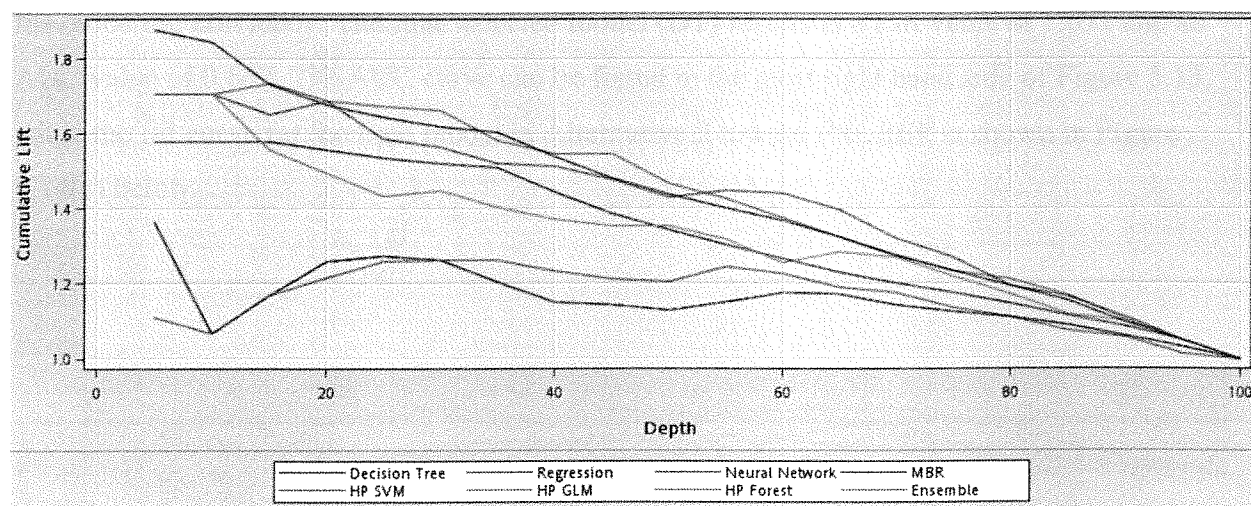


Figure 5.14(a) Lift Curve with Training Data Set toward PRCREV in Alternative 2

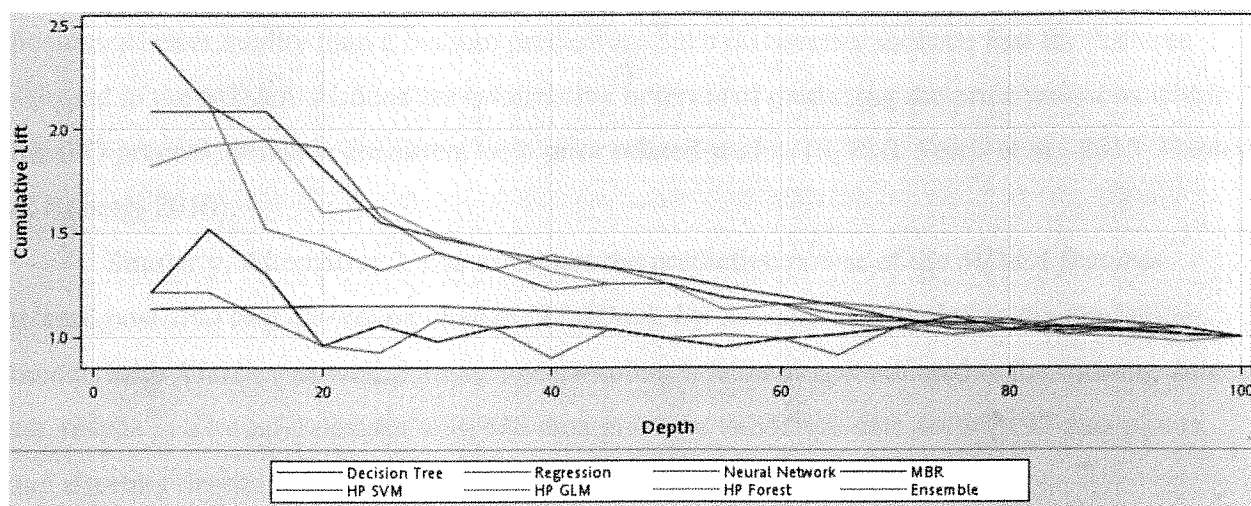


Figure 5.14(b) Lift Curve with Validation Data Set toward PRCREV in Alternative 2

With DT selected as the most accurate model, the testing date set is used to finally assess the modeling efficiency. The final selected model (DT) results in a Lift value of 1.671 and an AUC value of 0.689. The AUC curve can be found in the most right hand side of Figure 5.13, while the lift curve for the final model in Alternative 2 toward *PRCREV* is shown in Figure 5.14(c) below.

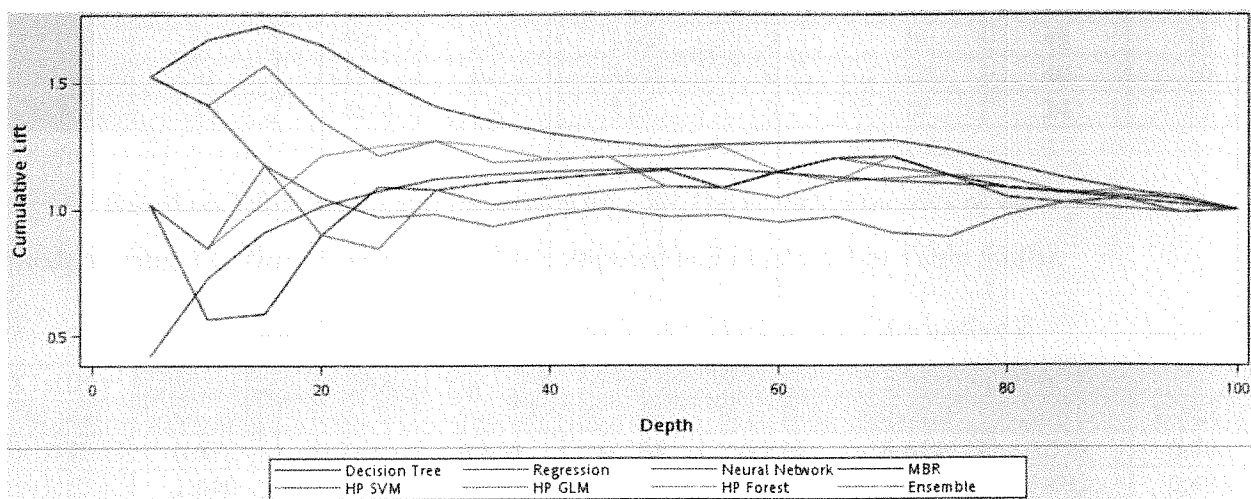


Figure 5.14(c) Lift Curve with Testing Data Set toward PRCREV in Alternative 2

Aforementioned efficiency metrics should be interpreted as follows: for predicting pre-IPO pricing revision using *MD&A features*, with the help of IPO characteristics, the prediction possesses better quality than a random prediction. Such discoveries indicate that the features selected in the MD&A sections are informative in terms of predicting the price revisions within the IPO process, which is consistent with prior related studies (S. P. S. Ferris et al., 2013; Hanley & Hoberg, 2010).

Similarly, Alternative 2 also examines the prediction power of the *MD&A features* toward post-IPO first day return change (*X1stDay*). *Up Revision* is included in the predictive models with *X1stDay* as well. Similar oversampling treatment is conducted to the training data set, results in a training data set with 643 data points, a validation data set with 92 data points, and a testing data set with 84 data points.

Table 5.6(a) shows the confusion matrices, which provide evidences for computing the accuracy metrics (precision, recall, accuracy, F-score).

Table 5.6(a). Confusion Matrices of X1stDay in Alternative 2

X1stDay	Training Data Set				Validation Data Set			
Model	FP	TP	FN	TN	FP	TP	FN	TN
DT	29	198	80	336	9	44	28	11
RF	0	100	178	365	4	21	51	16
EN	5	185	93	360	6	46	26	14
ANN	16	208	70	349	6	44	28	14
LR	5	158	120	360	2	29	43	18
GLM	5	158	120	360	2	29	43	18
SVM	0	137	141	365	0	32	40	20
k-NN	49	148	130	316	9	38	34	11

Based on Table 5.6(a) and Equations 8a – 8d, the accuracy metrics of each predictive models using *X1stDay* in Alternative 2 is computed and illustrated in Table 5.6(b).

Table 5.6(b). Accuracy Metrics of X1stDay in Alternative 2

X1stDay	Training Data Set				Validation Data Set			
Model	Precision	Recall	Accuracy	F-Score	Precision	Recall	Accuracy	F-Score
DT	0.872	0.712	0.830	0.784	0.830	0.611	0.598	0.704
RF	1.000	0.360	0.723	0.529	0.840	0.292	0.402	0.433
EN	0.974	0.665	0.848	0.791	0.885	0.639	0.652	0.742
ANN	0.929	0.748	0.866	0.829	0.880	0.611	0.630	0.721
LR	0.969	0.568	0.806	0.717	0.935	0.403	0.511	0.563
GLM	0.969	0.568	0.806	0.717	0.935	0.403	0.511	0.563
SVM	1.000	0.493	0.781	0.660	1.000	0.444	0.565	0.615
k-NN	0.751	0.532	0.722	0.623	0.809	0.528	0.533	0.639

The accuracy metrics of the predictive models with the training data set illustrate how the model is trained (shown on the left four columns of Table 5.6(b)); while the accuracy metrics of the predictive models with the validation data set (shown on the right four columns of Table 5.6(b)) demonstrate the accuracies of the trained models. As discussed in Section 4.4.3, F-score of the validation data set (most right column in Table 5.6(b)) is determined as the metric for selecting the predictive model for final assessment: thus, EN is selected as the most accurate model of Alternative 2 with *X1stDay* as the target variable. It is worth noting that the prediction toward *X1stDay* is comparable toward *PRCREV*, indicating that *MD&A features* are equivalently significant toward both pre- and post-IPO pricing trends.

On the other hand, the efficiency metrics (Lift and AUC) are also examined for the predictive models, with training and validation data sets. The results are reported in Table 5.6(c) below.

Table 5.6(c) Efficiency Metrics of X1stDay in Alternative 2

X1stDay	Training Data Set		Validation Data Set	
Model	Lift	AUC	Lift	AUC
DT	1.440	0.850	1.268	0.635
RF	1.760	0.950	0.958	0.499
EN	1.600	0.930	0.958	0.719
ANN	1.630	0.910	1.916	0.724
LR	1.460	0.840	2.874	0.680
GLM	1.460	0.840	2.874	0.679
SVM	1.270	0.820	1.916	0.673
k-NN	1.650	0.780	0.000	0.508

The AUC diagram of the predictive models with the training, validation, and testing data sets are shown in Figure 5.15 below (from left to right).

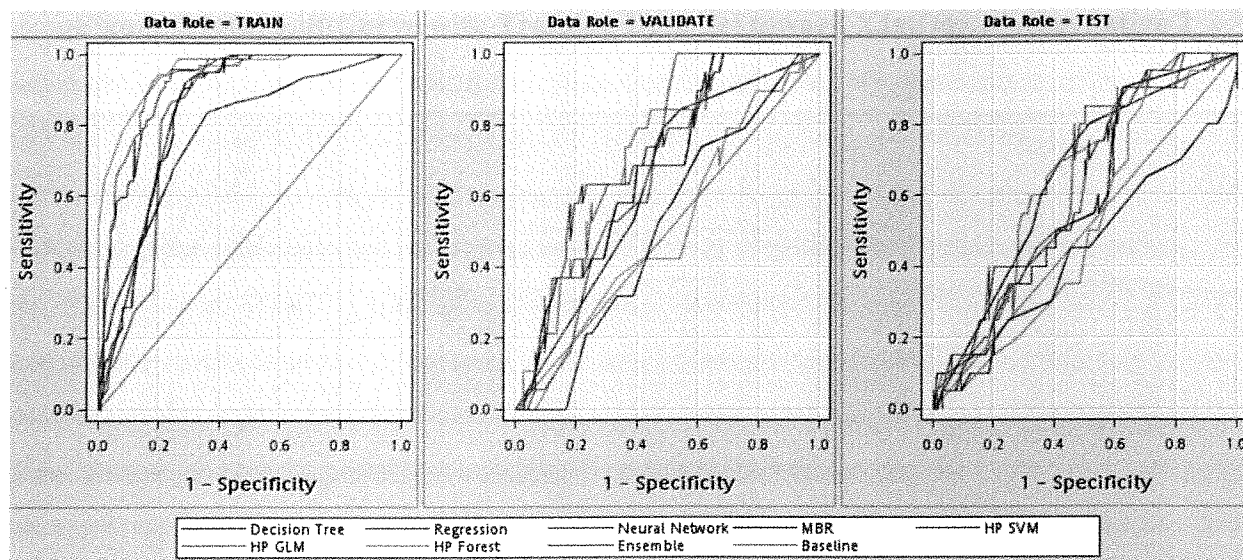


Figure 5.15. AUC Diagram of Predictive Models toward X1stDay in Alternative 2

The Lift Curves for predictive models in Alternative 2 toward *X1stDay* are shown in Figure 5.16(a) and (b) below.

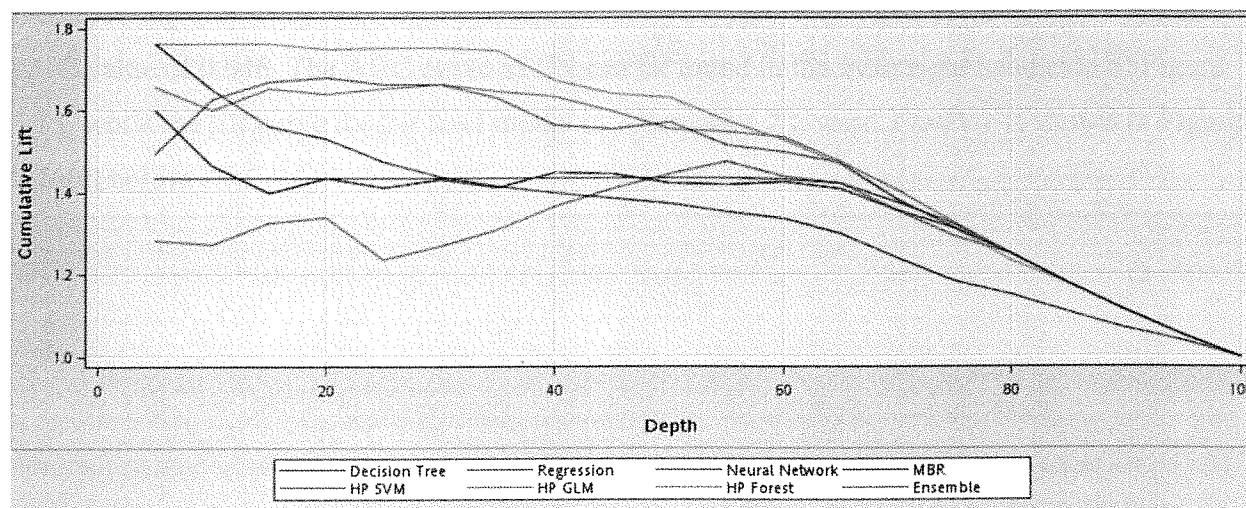


Figure 5.16(a) Lift Curve with Training Data Set toward X1stDay in Alternative 2

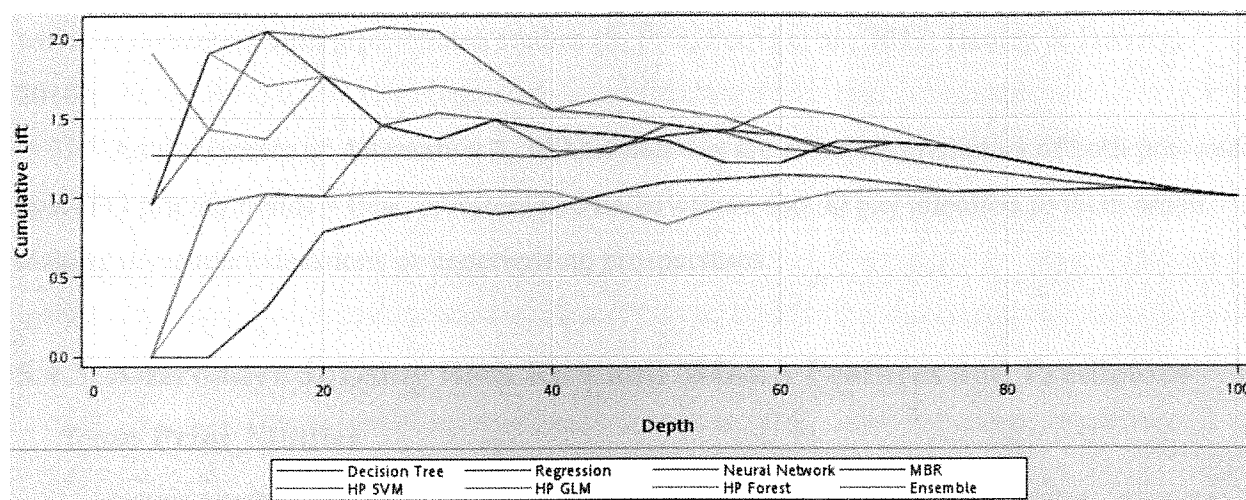


Figure 5.16(b) Lift Curve with Validation Data Set toward X1stDay in Alternative 2

With EN selected as the most accurate model, the testing date set is used to finally assess the modeling efficiency. The final selected model (EN) results in a Lift value of 1.972 and an AUC value of 0.646. The AUC curve of EN can be found in the most right hand side of Figure 5.15, while the lift curve for the final model in Alternative 2 toward *X1stDay* is shown in Figure 5.16(c) below.

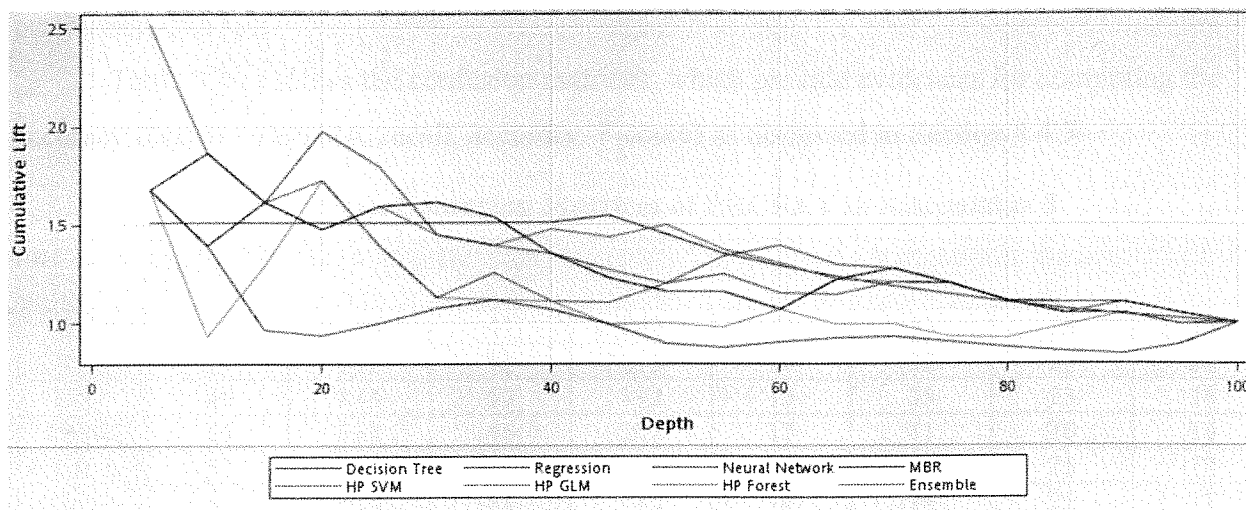


Figure 5.16(c) Lift Curve with Testing Data Set toward X1stDay in Alternative 2

Aforementioned efficiency metrics should be interpreted as follows: for predicting post-IPO pricing revision using *MD&A features*, with the help of IPO characteristics, the prediction

possesses better quality than a random prediction. Such discoveries indicate that the features selected in the *MD&A* sections are informative in terms of predicting the first day initial returns, which is consistent with prior related studies (S. P. S. Ferris et al., 2013; Hanley & Hoberg, 2010).

As a summary of Alternative 2, *MD&A features* are effective predictors of both pre- and post-IPO pricing trends – thus, investors and underwriters should pay attention to them when making investment decisions or underwriting prospectuses.

5.4.3. Alternative 3: Using Both Risk and MD&A Features and Predictors from Prior Studies

In this alternative, the predictive power of both the *risk-* and *MD&A features*, in combination with the predictors from prior studies, toward pre- and post-IPO pricing trends, is examined.

Firstly, pre-IPO price revision (*PRCREV*) is selected as the target variable in the predictive model. Both *risk-features* and *MD&A feature* predictors are also selected, along with the predictors from prior studies – excluding *Up Revision*. Similar oversampling treatment is conducted, as described in previous alternatives, yielding same resulting data points in the partitioned data sets.

Table 5.7(a) shows the confusion matrices, which provide evidences for computing the accuracy metrics (precision, recall, accuracy, F-score) as discussed in Section 4.4.3.

Table 5.7(a). Confusion Matrices of PRCREV in Alternative 3

PRCREV	Training Data Set				Validation Data Set			
Model	FP	TP	FN	TN	FP	TP	FN	TN
DT	90	157	47	142	17	43	18	14
RF	19	112	92	213	10	32	29	20
EN	40	145	59	192	15	38	23	16
ANN	25	158	46	207	17	41	20	14
LR	57	121	83	175	14	36	25	17
GLM	57	121	83	175	14	36	25	17
SVM	40	109	95	192	8	35	26	23
k-NN	36	123	81	196	12	29	32	19

Based on Table 5.7(a) and Equations 8a – 8d, the accuracy metrics of each predictive models using *PRCREV* in Alternative 3 is computed and illustrated in Table 5.7(b).

Table 5.7(b). Accuracy Metrics of PRCREV in Alternative 3

PRCREV	Training Data Set				Validation Data Set			
Model	Precision	Recall	Accuracy	F-Score	Precision	Recall	Accuracy	F-Score
DT	0.636	0.770	0.686	0.696	0.717	0.705	0.620	0.711
RF	0.855	0.549	0.745	0.669	0.762	0.525	0.571	0.621
EN	0.784	0.711	0.773	0.746	0.717	0.623	0.587	0.717
ANN	0.863	0.775	0.837	0.817	0.707	0.672	0.598	0.689
LR	0.680	0.593	0.679	0.634	0.720	0.590	0.576	0.649
GLM	0.680	0.593	0.679	0.634	0.720	0.590	0.576	0.649
SVM	0.732	0.534	0.690	0.618	0.814	0.574	0.630	0.673
k-NN	0.774	0.603	0.732	0.678	0.707	0.475	0.522	0.569

The accuracy metrics of the predictive models with the training data set illustrate how the model is trained (shown on the left four columns of Table 5.7(b)); while the accuracy metrics of the predictive models with the validation data set (shown on the right four columns of Table 5.7(b)) demonstrate the accuracies of the trained models. Similarly, F-score of the validation data set (most right column in Table 5.7(b)) is determined as the metric for selecting the predictive model for final assessment: thus, EN is selected as the most accurate model of Alternative 3 with *PRCREV* as the target variable.

On the other hand, the efficiency metrics (Lift and AUC) are also examined for the predictive models, with training and validation data sets. The results are reported in Table 5.7(c) below. Values of these metrics also support the selection of EN as the final model in predicting *PRCREV* with both *risk-* and *MD&A features*.

Table 5.7(c) Efficiency Metrics of PRCREV in Alternative 3

PRCREV	Training Data Set		Validation Data Set	
Model	Lift	AUC	Lift	AUC
DT	1.620	0.809	1.625	0.653
RF	1.673	0.820	1.820	0.583
EN	1.879	0.901	1.820	0.668
ANN	1.751	0.892	1.213	0.649
LR	1.623	0.740	1.517	0.613
GLM	1.623	0.740	1.213	0.613
SVM	1.281	0.733	1.213	0.627
k-NN	1.862	0.825	1.075	0.523

The AUC diagram of the predictive models with the training, validation, and testing data sets are shown in Figure 5.17 below (from left to right).

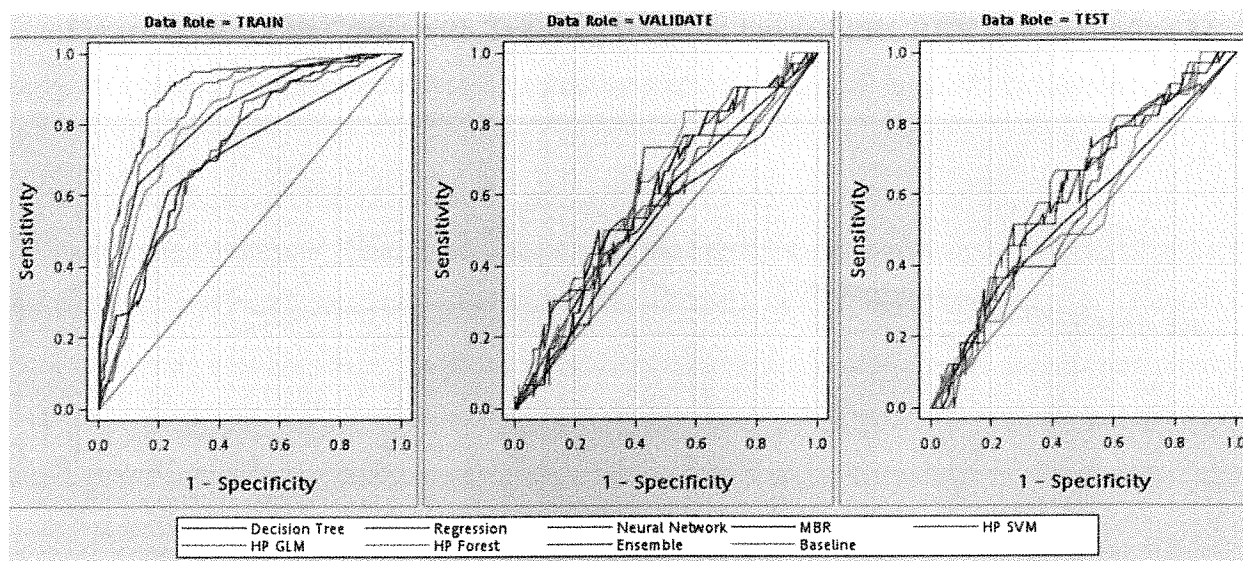


Figure 5.17. AUC Diagram of Predictive Models toward PRCREV in Alternative 3

The Lift Curves for predictive models in Alternative 3 toward *PRCREV* with training and validation data set are shown in Figure 5.18 (a) and (b) below, respectively.

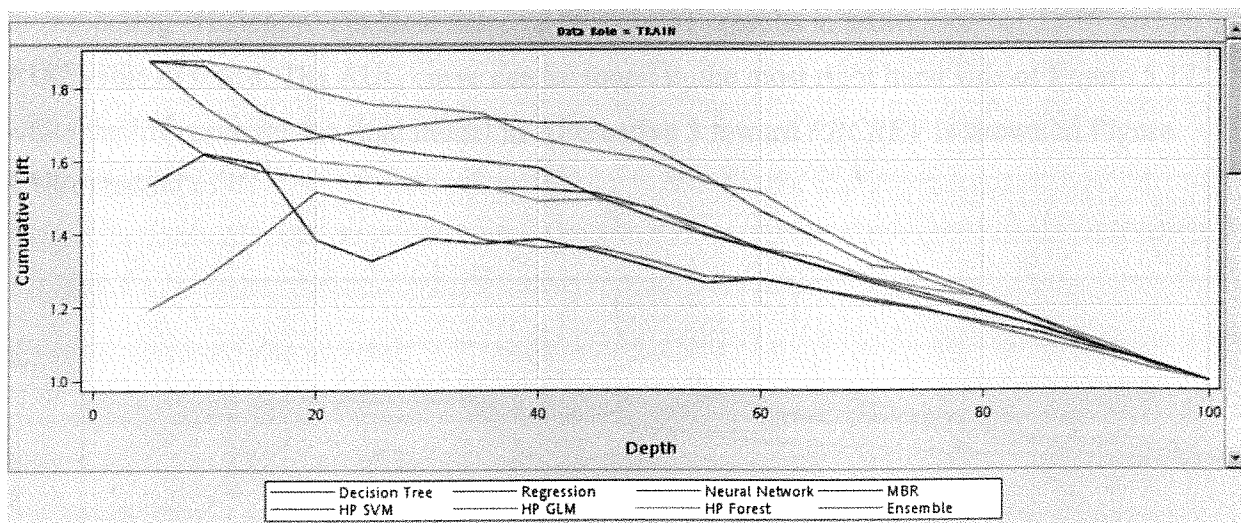


Figure 5.18(a) Lift Curve with Training Data Set toward PRCREV in Alternative 3

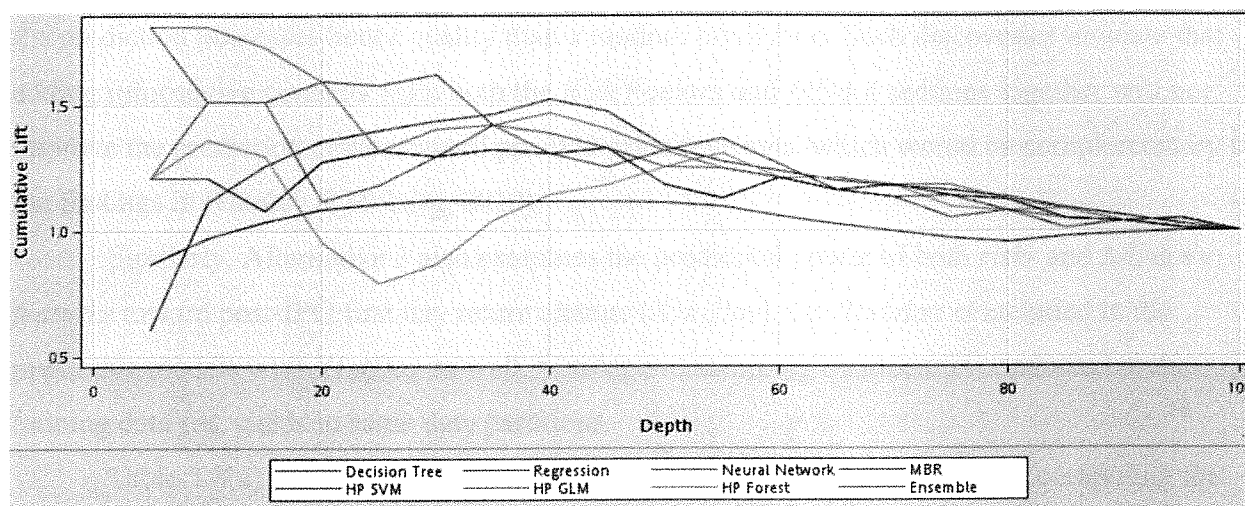


Figure 5.18(b) Lift Curve with Validation Data Set toward PRCREV in Alternative 3

With DT selected as the most accurate model, the testing data set is used to finally assess the modeling efficiency. The final selected model (DT) results in a Lift value of 1.909 and an AUC value of 0.614. The AUC curve can be found in the most right hand side of Figure 5.17, while the lift curve for the final model in Alternative 3 toward *PRCREV* is shown in Figure 5.18(c) below.

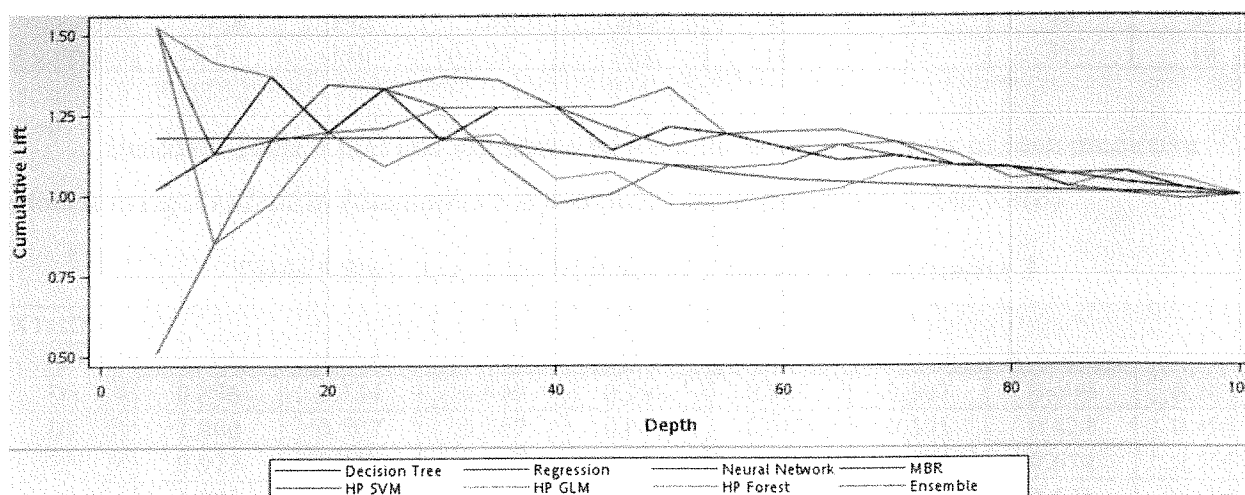


Figure 5.18(c) Lift Curve with Testing Data Set toward PRCREV in Alternative 3

Aforementioned efficiency metrics should be interpreted as follows: for predicting pre-IPO pricing revision using both *risk-* and *MD&A features*, with the help of IPO characteristics, the prediction possesses better quality than a random prediction. Such discoveries indicate that adding informative contents from both the *Risk Factors* and *MD&A* sections together will not improve the prediction power toward pre-IPO price revisions, which would be complementary to the findings in (Hanley & Hoberg, 2010).

Similarly, Alternative 3 also examines the prediction power of both *risk-* and *MD&A features* toward post-IPO first day return change (*X1stDay*). *Up Revision* is included in the predictive models with *X1stDay* as well. Similar oversampling treatment is conducted to the training data set, yields in same data partitions.

Table 5.8(a) shows the confusion matrices, which provide evidences for computing the accuracy metrics (precision, recall, accuracy, F-score).

Table 5.8(a). Confusion Matrices of X1stDay in Alternative 3

X1stDay	Training Data Set				Validation Data Set			
Model	FP	TP	FN	TN	FP	TP	FN	TN
DT	29	204	74	336	7	43	29	13
RF	0	141	137	365	7	26	46	13
EN	5	214	64	360	7	49	23	13
ANN	0	230	48	365	8	41	31	12
LR	0	193	85	365	8	36	36	12
GLM	0	193	85	365	8	36	36	12
SVM	0	143	135	365	0	32	40	20
k-NN	47	150	128	318	10	35	37	10

Based on Table 5.8(a) and Equations 8a – 8d, the accuracy metrics of each predictive models using *X1stDay* in Alternative 3 is computed and illustrated in Table 5.8(b).

Table 5.8(b). Accuracy Metrics of X1stDay in Alternative 3

X1stDay	Training Data Set				Validation Data Set			
Model	Precision	Recall	Accuracy	F-Score	Precision	Recall	Accuracy	F-Score
DT	0.876	0.734	0.840	0.798	0.860	0.597	0.609	0.705
RF	1.000	0.507	0.787	0.673	0.788	0.361	0.424	0.495
EN	0.977	0.770	0.893	0.861	0.875	0.681	0.674	0.766
ANN	1.000	0.827	0.925	0.906	0.837	0.569	0.576	0.678
LR	1.000	0.694	0.868	0.820	0.818	0.500	0.522	0.621
GLM	1.000	0.694	0.868	0.820	0.818	0.500	0.522	0.621
SVM	1.000	0.514	0.790	0.679	1.000	0.444	0.565	0.775
k-NN	0.761	0.540	0.728	0.632	0.778	0.486	0.489	0.598

The accuracy metrics of the predictive models with the training data set illustrate how the model is trained (shown on the left four columns of Table 5.8(b)); while the accuracy metrics of the predictive models with the validation data set (shown on the right four columns of Table 5.8(b)) demonstrate the accuracies of the trained models. As discussed above, F-score of the validation data set (most right column in Table 5.8(b)) is determined as the metric for selecting the predictive model for final assessment: thus, SVM is selected as the most accurate model of Alternative 3 with *X1stDay* as the target variable..

On the other hand, the efficiency metrics (Lift and AUC) are also examined for the predictive models, with training and validation data sets. The results are reported in Table 5.8(c) below.

Table 5.8(c) Efficiency Metrics of X1stDay in Alternative 3

X1stDay	Training Data Set		Validation Data Set	
	Lift	AUC	Lift	AUC
DT	1.450	0.860	1.268	0.671
RF	1.760	0.970	0.479	0.472
EN	1.760	0.960	2.395	0.700
ANN	1.710	0.950	1.916	0.691
LR	1.570	0.880	1.916	0.668
GLM	1.570	0.880	1.916	0.668
SVM	1.540	0.850	0.958	0.725
k-NN	1.640	0.770	0.000	0.486

The AUC diagram of the predictive models with the training, validation, and testing data sets are shown in Figure 5.19 below (from left to right).

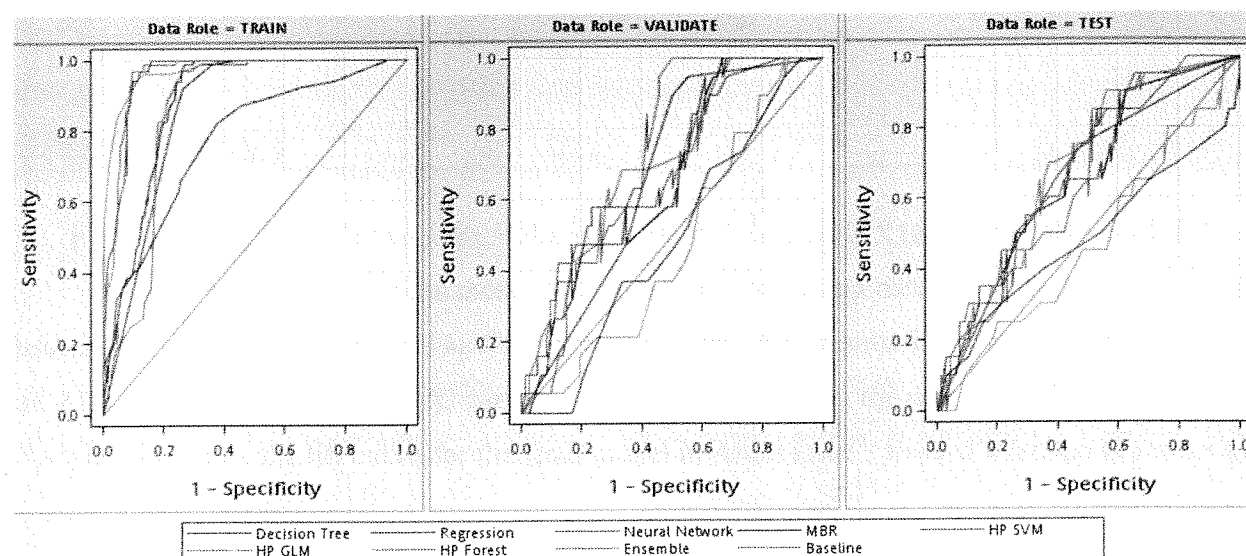


Figure 5.19. AUC Diagram of Predictive Models toward X1stDay in Alternative 3

The Lift Curves for predictive models in Alternative 3 toward *X1stDay* are shown in Figure 5.20(a) and (b) below.

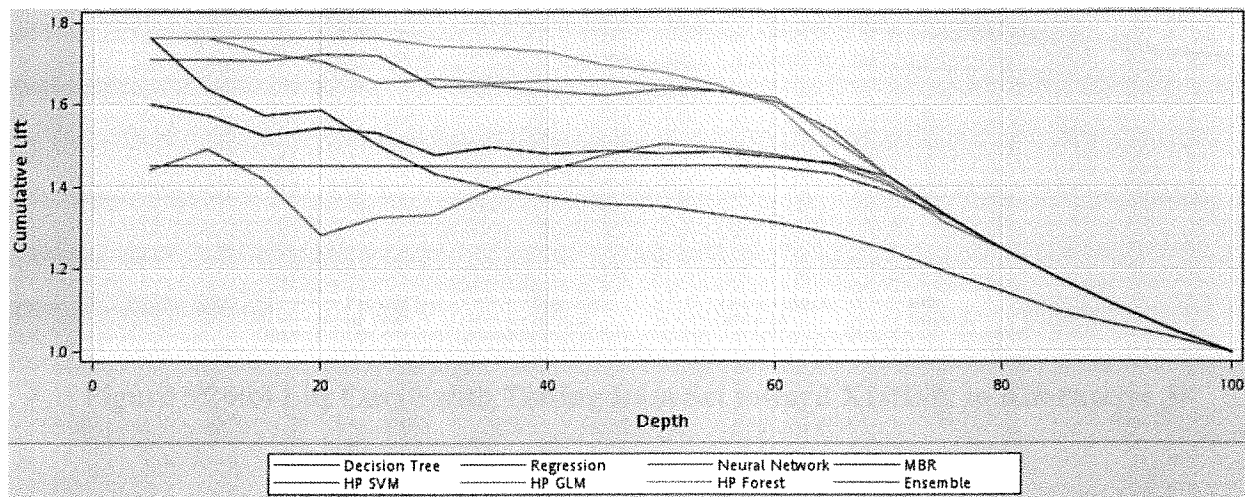


Figure 5.20(a) Lift Curve with Training Data Set toward X1stDay in Alternative 3

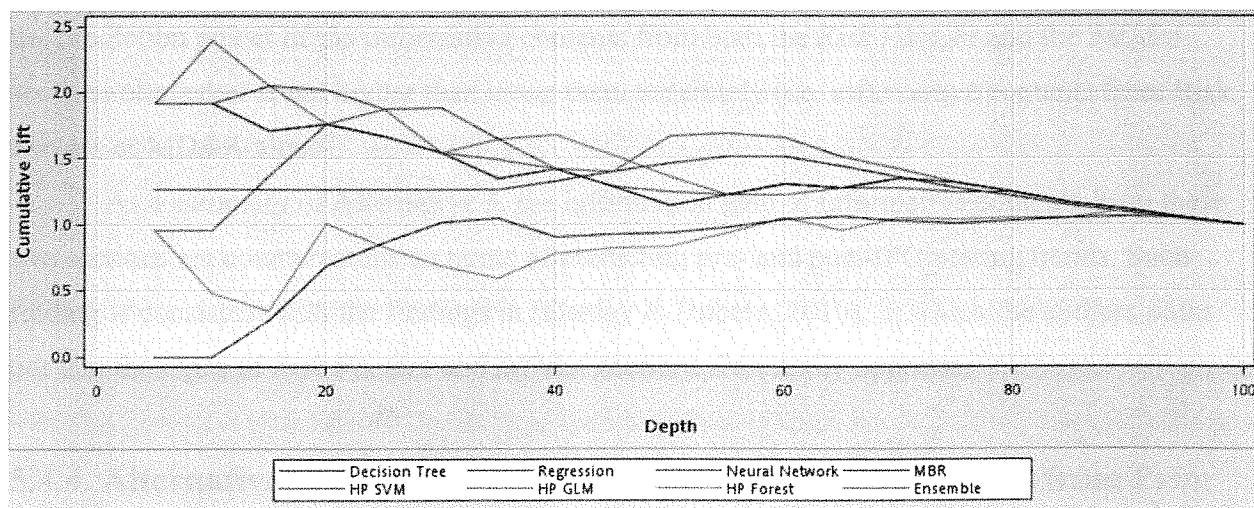


Figure 5.20(b) Lift Curve with Validation Data Set toward X1stDay in Alternative 3

With SVM selected as the most accurate model, the testing date set is used to finally assess the modeling efficiency. The final selected model (EN) results in a Lift value of 1.867 and an AUC value of 0.648. The AUC curve of SVM can be found in the most right hand side of Figure 5.19, while the lift curve for the final model in Alternative 3 toward *X1stDay* is shown in Figure 5.20(c) below.

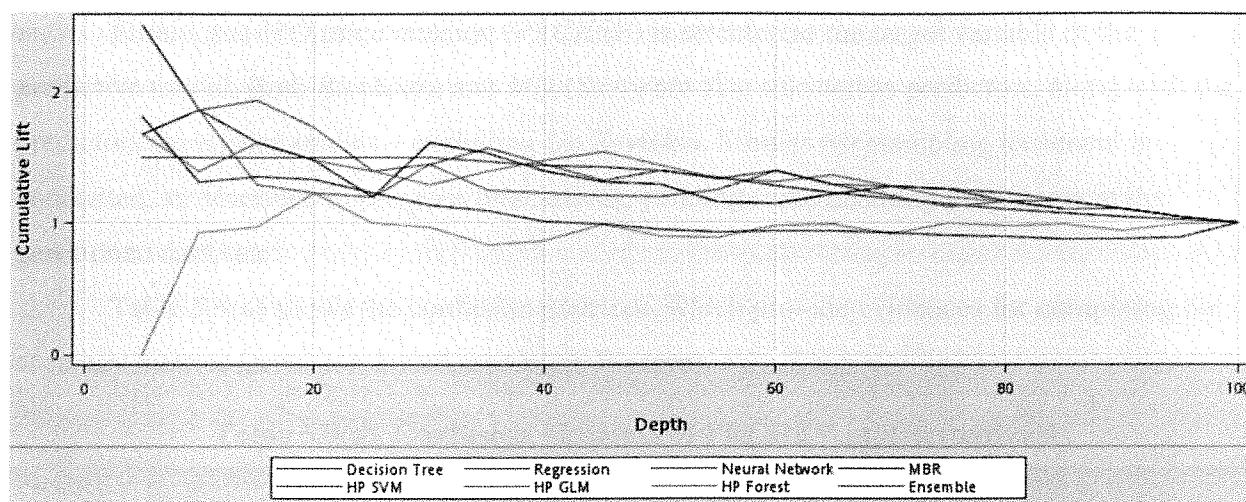


Figure 5.20(c) Lift Curve with Testing Data Set toward X1stDay in Alternative 3

Aforementioned efficiency metrics should be interpreted as follows: for predicting post-IPO pricing revision using both *risk*- and *MD&A features*, with the help of IPO characteristics, the prediction possesses better quality than a random prediction. Such discoveries indicate that the prediction power of the informative contents from both the *Risk Factors* and the *MD&A* sections altogether is not higher than using them separately (i.e. informative contents from Risk Factors or MD&A alone).

As a summary of Alternative 3, the information gain of informative contents from these two sections are counteracting, in terms of predicting pre- and post-IPO pricing trends. Such finding is consistent with the findings in (Hanley & Hoberg, 2010), in which the authors point out that the tones of *Risk Factors* and *MD&A* sections are usually opposite.

5.4.4. Alternative 4: Using Aggregated Predictors and Predictors from Prior Studies

As discussed in Section 4.4.3, introducing too many variables in the predictive models may lead to the ‘*overfitting*’ issue. To mitigate this issue, we defined aggregated predictors (*RiskScore* and *MDAScore*) via equation (6) and (7) from Section 4.4.3, toward pre- and post-IPO pricing trends. In Alternative 4, these predictive models are examined using aforementioned metrics along similar process.

Firstly, pre-IPO price revision (*PRCREV*) is selected as the target variable in the predictive model. Both *RiskScore* and *MDAScore* are also selected as predictors, along with the predictors from prior studies – excluding *Up Revision*. Similar oversampling treatment is conducted, as described in previous alternatives, yielding same resulting data points in the partitioned data sets.

Table 5.9(a) shows the confusion matrices, which provide evidences for computing the accuracy metrics (precision, recall, accuracy, F-score).

Table 5.9(a). Confusion Matrices of PRCREV in Alternative 4

PRCREV	Training Data Set				Validation Data Set			
Model	FP	TP	FN	TN	FP	TP	FN	TN
DT	84	162	42	148	16	42	19	15
RF	29	98	106	203	11	28	33	19
EN	54	143	61	178	13	33	28	18
ANN	47	132	72	185	14	29	32	17
LR	41	90	114	191	10	24	37	21
GLM	41	90	114	191	10	24	37	21
SVM	22	51	153	210	8	18	43	23
k-NN	35	123	81	197	12	29	32	19

Based on Table 5.7(a) and Equations 8a – 8d, the accuracy metrics of each predictive models using *PRCREV* in Alternative 4 is computed and illustrated in Table 5.9(b).

Table 5.9(b). Accuracy Metrics of PRCREV in Alternative 4

PRCREV	Training Data Set				Validation Data Set			
Model	Precision	Recall	Accuracy	F-Score	Precision	Recall	Accuracy	F-Score
DT	0.659	0.794	0.711	0.720	0.724	0.689	0.620	0.706
RF	0.772	0.480	0.690	0.592	0.718	0.459	0.516	0.560
EN	0.726	0.701	0.736	0.713	0.717	0.541	0.554	0.617
ANN	0.737	0.647	0.727	0.689	0.674	0.475	0.500	0.558
LR	0.687	0.441	0.644	0.537	0.706	0.393	0.489	0.505
GLM	0.687	0.441	0.644	0.537	0.706	0.393	0.489	0.505
SVM	0.699	0.250	0.599	0.368	0.692	0.295	0.446	0.414
k-NN	0.778	0.603	0.734	0.680	0.727	0.475	0.622	0.769

The accuracy metrics of the predictive models with the training data set illustrate how the model is trained (shown on the left four columns of Table 5.9(b)); while the accuracy metrics of the predictive models with the validation data set (shown on the right four columns of Table 5.9(b)) demonstrate the accuracies of the trained models. Similarly, F-score of the validation data set (most right column in Table 5.9(b)) is determined as the metric for selecting the predictive

model for final assessment: thus, k-NN is selected as the most accurate model of Alternative 4 with *PRCREV* as the target variable.

On the other hand, the efficiency metrics (Lift and AUC) are also examined for the predictive models, with training and validation data sets. The results are reported in Table 5.9(c) below. Values of these metrics also support the selection of k-NN as the final model in predicting *PRCREV* with aggregated predictors.

Table 5.9(c) Efficiency Metrics of PRCREV in Alternative 4

PRCREV	Training Data Set		Validation Data Set	
	Lift	AUC	Lift	AUC
DT	1.532	0.737	1.321	0.569
RF	1.554	0.776	1.517	0.578
EN	1.708	0.794	1.517	0.578
ANN	1.623	0.775	1.213	0.549
LR	1.111	0.633	1.517	0.567
GLM	1.111	0.633	1.517	0.567
SVM	0.897	0.613	1.517	0.561
k-NN	1.484	0.705	1.618	0.741

The AUC diagram of the predictive models with the training, validation, and testing data sets are shown in Figure 5.21 below (from left to right).

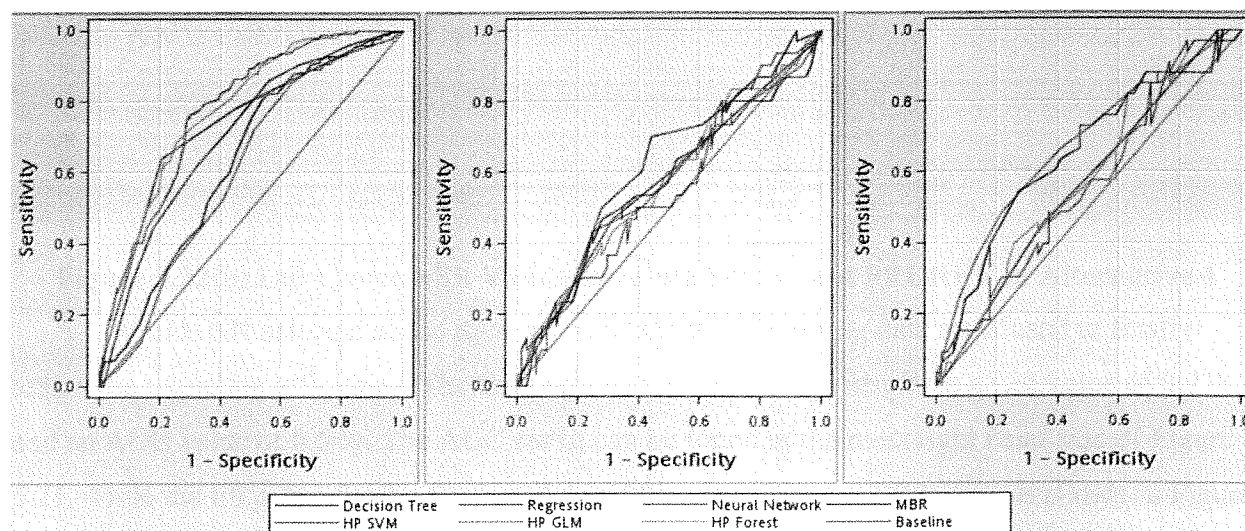


Figure 5.21. AUC Diagram of Predictive Models toward PRCREV in Alternative 4

The Lift Curves for predictive models in Alternative 4 toward *PRCREV* with training and validation data set are shown in Figure 5.22 (a) and (b) below, respectively.

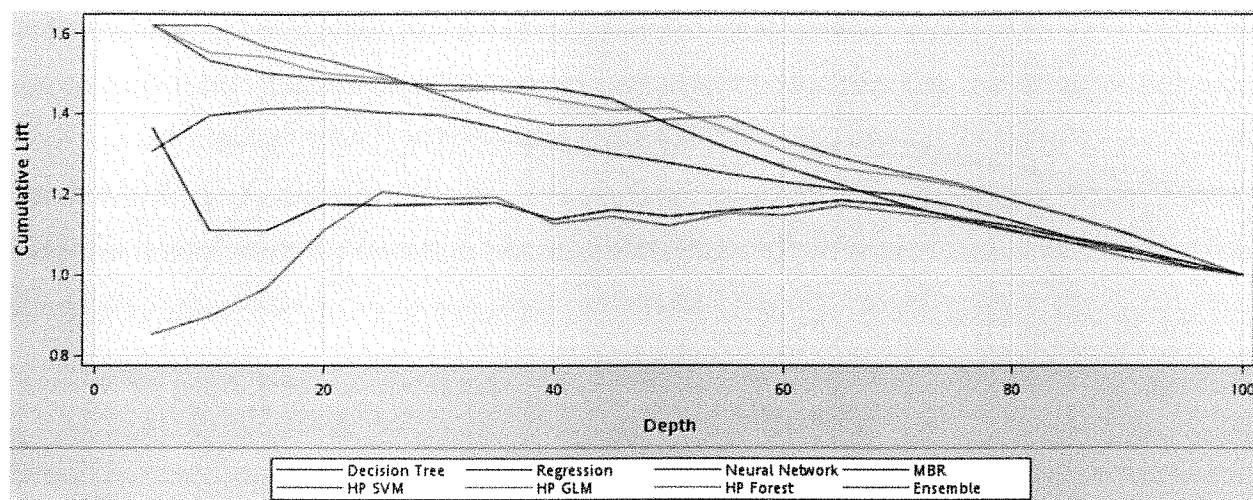


Figure 5.22(a) Lift Curve with Training Data Set toward PRCREV in Alternative 4

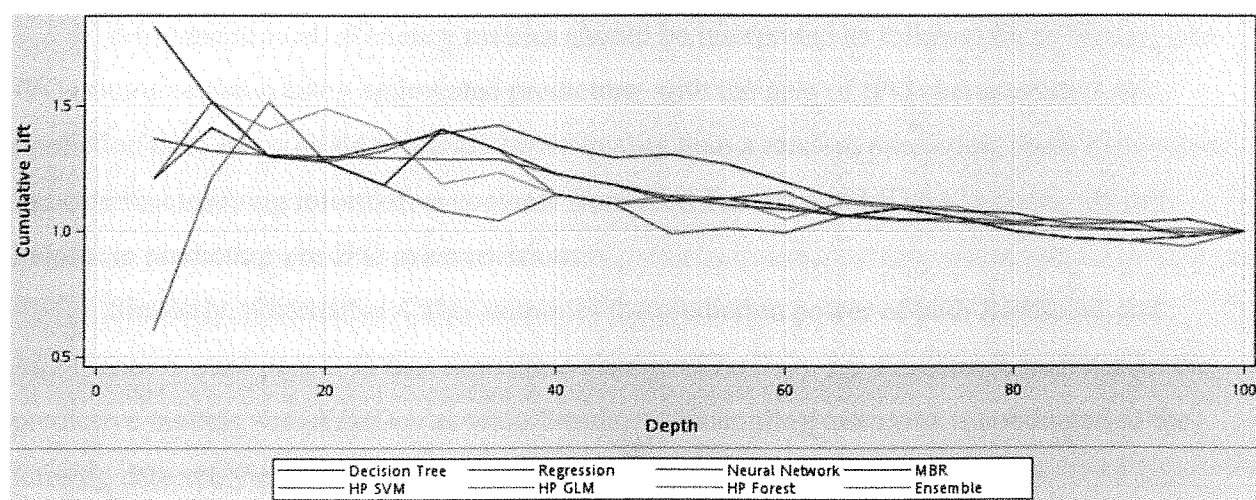


Figure 5.22(b) Lift Curve with Validation Data Set toward PRCREV in Alternative 4

With k-NN selected as the most accurate model, the testing date set is used to finally assess the modeling efficiency. The final selected model (k-NN) results in a Lift value of 0.978 and an AUC value of 0.544. The AUC curve can be found in the most right hand side of Figure 5.21, while the lift curve for the final model in Alternative 4 toward *PRCREV* is shown in Figure 5.22(c) below.

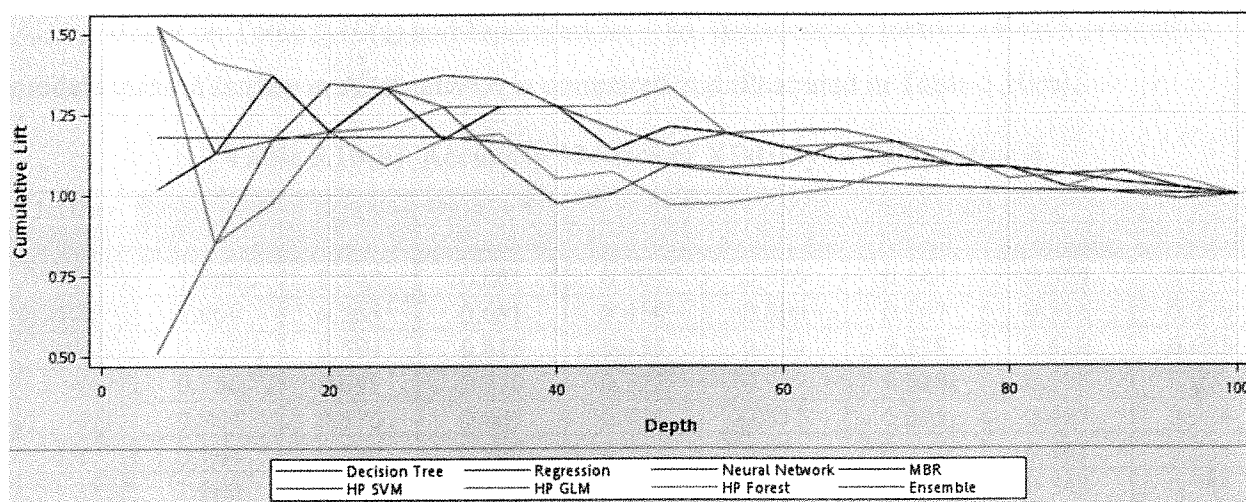


Figure 5.22(c) Lift Curve with Testing Data Set toward PRCREV in Alternative 3

Aforementioned efficiency metrics should be interpreted as follows: for predicting pre-IPO pricing revision using aggregated predictors, with the help of IPO characteristics, the prediction does not provide evidently better quality than a random prediction. Such discoveries indicate that unifying informative contents from Risk Factors and MD&A sections are not helpful in predicting pre-IPO price revisions.

Similarly, Alternative 4 also examines the prediction power of both *RiskScore* and *MDAScore* toward post-IPO first day return change (*X1stDay*). *Up Revision* is included in the predictive models with *X1stDay* as well. Similar oversampling treatment is conducted to the training data set, yields in same data partitions.

Table 5.10(a) shows the confusion matrices, which provide evidences for computing the accuracy metrics (precision, recall, accuracy, F-score).

Table 5.10(a). Confusion Matrices of *X1stDay* in Alternative 4

<i>X1stDay</i>	Training Data Set				Validation Data Set			
Model	FP	TP	FN	TN	FP	TP	FN	TN
DT	0	150	128	365	1	32	40	19
RF	0	73	205	365	3	12	60	17
EN	6	165	113	359	1	38	34	19
ANN	62	192	86	303	3	38	34	17
LR	6	160	118	359	4	27	45	16
GLM	6	160	118	359	4	27	45	16
SVM	0	136	142	365	0	30	42	20
k-NN	49	153	125	316	10	37	35	10

Based on Table 5.10(a) and Equations 8a – 8d, the accuracy metrics of each predictive models using *X1stDay* in Alternative 4 is computed and illustrated in Table 5.10(b).

Table 5.10(b). Accuracy Metrics of X1stDay in Alternative 4

X1stDay	Training Data Set				Validation Data Set			
Model	Precision	Recall	Accuracy	F-Score	Precision	Recall	Accuracy	F-Score
DT	1.000	0.540	0.801	0.701	0.970	0.444	0.554	0.610
RF	1.000	0.263	0.681	0.416	0.800	0.167	0.315	0.276
EN	0.965	0.594	0.815	0.735	0.974	0.528	0.620	0.685
ANN	0.756	0.691	0.770	0.722	0.927	0.528	0.598	0.673
LR	0.964	0.576	0.807	0.721	0.871	0.375	0.467	0.524
GLM	0.964	0.576	0.807	0.721	0.871	0.375	0.467	0.524
SVM	1.000	0.489	0.779	0.657	1.000	0.417	0.543	0.588
k-NN	0.757	0.550	0.729	0.638	0.787	0.514	0.511	0.622

The accuracy metrics of the predictive models with the training data set illustrate how the model is trained (shown on the left four columns of Table 5.10(b)); while the accuracy metrics of the predictive models with the validation data set (shown on the right four columns of Table 5.10(b)) demonstrate the accuracies of the trained models. As discussed above, F-score of the validation data set (most right column in Table 5.10(b)) is determined as the metric for selecting the predictive model for final assessment: thus, EN is selected as the most accurate model of Alternative 10 with *X1stDay* as the target variable..

On the other hand, the efficiency metrics (Lift and AUC) are also examined for the predictive models, with training and validation data sets. The results are reported in Table 5.10(c) below. The values in Table 5.10 (c) also approve that the selection of EN as the final model (highest validation AUC of 0.713).

Table 5.10(c) Efficiency Metrics of X1stDay in Alternative 4

X1stDay	Training Data Set		Validation Data Set	
Model	Lift	AUC	Lift	AUC
DT	1.300	0.770	1.486	0.696
RF	1.660	0.840	0.599	0.497
EN	1.330	0.840	0.958	0.713
ANN	1.490	0.800	1.437	0.711
LR	1.460	0.820	0.958	0.625
GLM	1.460	0.820	0.958	0.625
SVM	1.270	0.800	1.437	0.640
k-NN	1.620	0.770	0.000	0.444

The AUC diagram of the predictive models with the training, validation, and testing data sets are shown in Figure 5.23 below (from left to right).

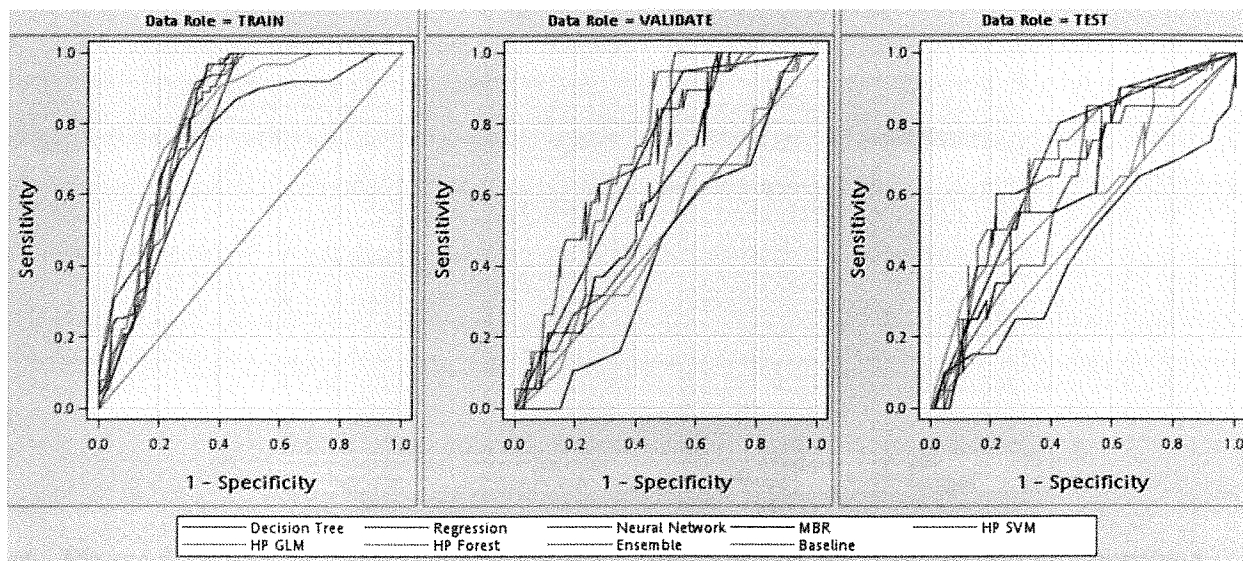


Figure 5.23. AUC Diagram of Predictive Models toward *X1stDay* in Alternative 4

The Lift Curves for predictive models in Alternative 4 toward *X1stDay* are shown in Figure 5.24(a) and (b) below.

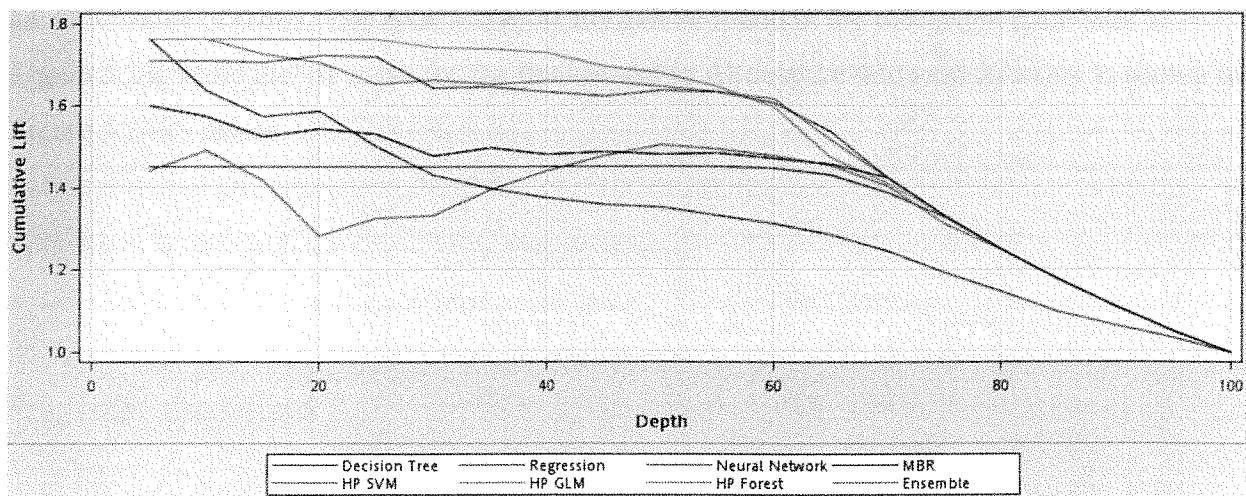


Figure 5.24(a) Lift Curve with Training Data Set toward X1stDay in Alternative 4

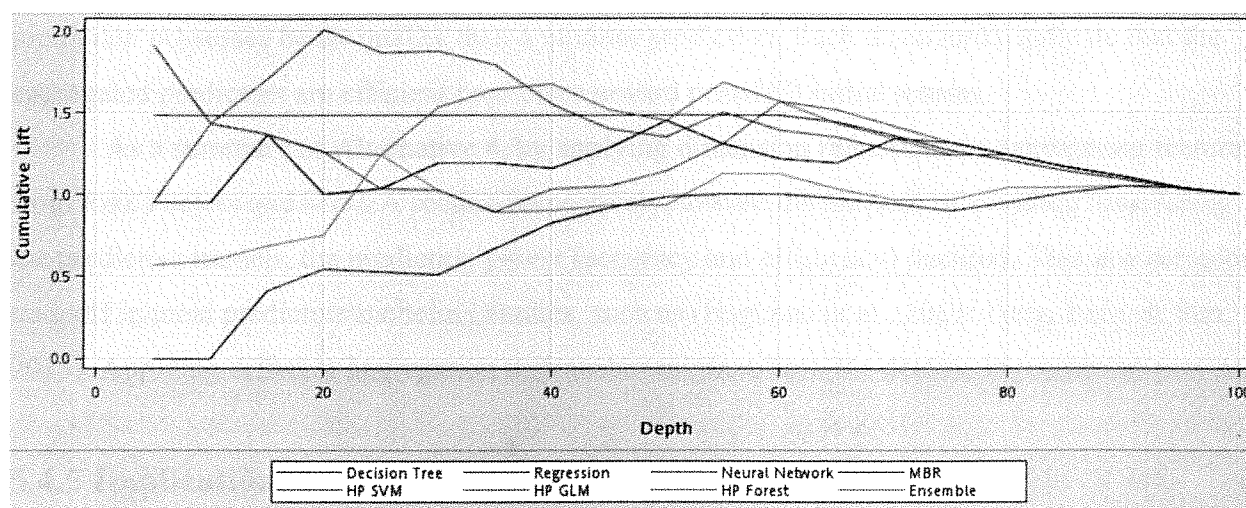


Figure 5.24(b) Lift Curve with Validation Data Set toward X1stDay in Alternative 4

With SVM selected as the most accurate model, the testing date set is used to finally assess the modeling efficiency. The final selected model (EN) results in a Lift value of 1.400 and an AUC value of 0.683. The AUC curve of EN can be found in the most right hand side of Figure 5.23, while the lift curve for the final model in Alternative 4 toward *X1stDay* is shown in Figure 5.24(c) below.

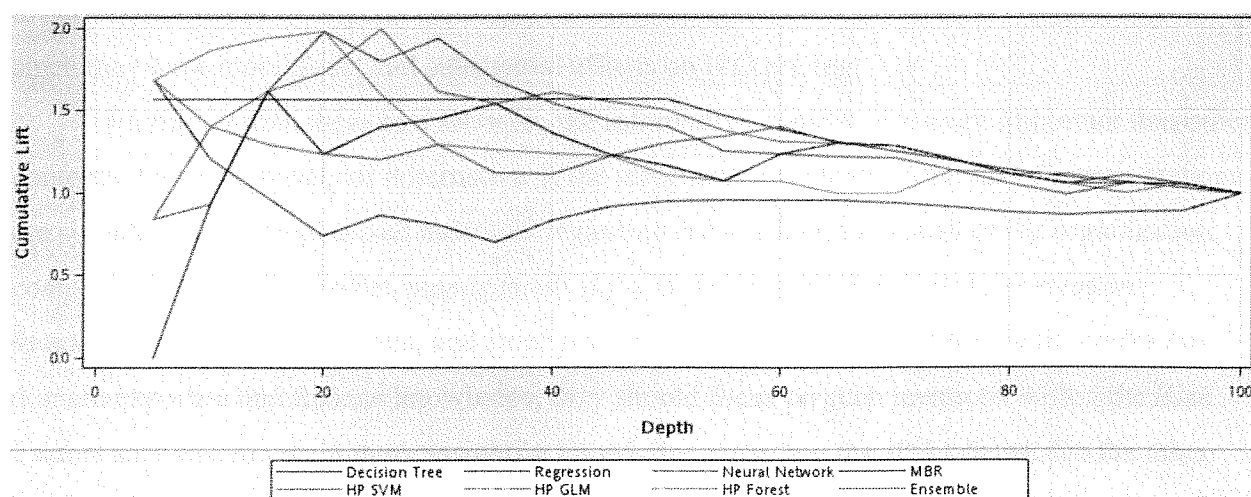


Figure 5.24(c) Lift Curve with Testing Data Set toward X1stDay in Alternative 4

Aforementioned efficiency metrics should be interpreted as follows: for predicting post-IPO pricing revision using aggregated predictors, with the help of IPO characteristics, the prediction possesses better quality than a random prediction. Such discoveries indicate that the aggregated predictors are efficient predictors toward post-IPO initial returns.

As a summary of Alternative 4, by applying dimension reduction via aggregating features from Risk Factors and MD&A section respectively, toward the purpose of avoiding ‘*overfitting*’ the predictive models, the prediction power (accuracy and efficiency) declines. This is a common tradeoff in most predictive modeling studies, such as (Hagenau et al., 2013; Tang, Liao, & Sun, 2013). Applying different aggregation approaches might relieve the decline in prediction power.

5.4.5 Implications to Practitioners

In this subsection, the findings from the experiments are summarized aligning with the motivations of this study.

Firstly, the experiment results verified that the *Risk Factors* and *MD&A* sections are more significant sections in IPO prospectuses, than other sections (such as *Prospectus Summary* and *Use of Proceeds*). Thus, whilst reading the IPO prospectuses, potential investors should focus on the contents of these two sections, in contrast to other sections in the prospectuses, to understand the pricing trends within and right after the IPO process. On the other hand, less experienced underwriters should also spend their time on fine-tuning the contents in these two sections – since they have more direct and substantial effects on IPO pricing.

Further, within these two sections, the informative contents are more important than other contents. From the results of Alternative 1, the informative contents in the *Risk Factors* sections are signaled by the emphasized mentions regarding risks related to *growth* of the organization, *competitiveness* in the market, *management skills* in the organization, *customer relations*, *lawsuits* and other legal matters, and *stock prices*. On one hand, for investors with *inadequate* domain expertise and finance knowledge, they should focus on such mentions within the Risk Factors sections, to obtain more insightful knowledge regarding the IPO pricing. On the other hand, underwriters should provide *more* informative contents regarding these features, in order to convince the investors to be more confident of the issuance.

Similarly, from the results of Alternative 2, the informative contents in the *MD&A* sections are *forward-looking*, *sentiment-subjective* sentences discussing *sales/revenues*,

costs/losses, investments, net incomes, and cash flows of the issuing organization. *Under-experienced* investor should spend more time reading these statements, since they are a direct proxy of the management's confidence of the future prospect. Investors should also consider disclose more information regarding these contents, in a *forward-looking, sentiment-subjective* manner while writing the *MD&A* sections.

There are a few other findings extracted from the experiments. For instance, from the comparison of Alternative 1 and 2, it is justified to claim that the *Risk Factors* sections are more informative (comparing to the *MD&A* sections), with respect to predict post-IPO pricings; while the *MD&A* features are more informative toward pre-IPO pricing trends. Such finding provides a direction of preference of reading and writing these two sections. Also, it is concluded in the experiment that the *pre-IPO price revisions* have a substantial effect on the *post-IPO initial returns*, which would be beneficial for potential investors.

CHAPTER 6

CONCLUSIONS

In this chapter, section 6.1 provides concluding remarks, summarizing the research study. Section 6.2 revisits the contributions of this work by evaluating the design artifacts against the design requirements presented earlier in Section 2.4. The relevance of this work is discussed in Section 6.3 along with recommendations to practitioners (investors and underwriters). The limitations and future steps of this work are discussed in Section 6.4.

6.1 Concluding Remarks

IPO prospectus is one of the most reliable sources from the point of view of exchanging crucial information between the investors and the issuers/underwriters and thus it holds significant value for knowledge discovery purposes and understanding the phenomena of the IPO process. Such information is disclosed via the issuance of the prospectus, which contains voluminous textual information across different versions.

As discussed in Section 1.4, there are two key research components of this project, namely *IPO ontology enrichment* and *IPO pricing predictions*. In terms of IPO ontology enrichment, we present an approach for enriching domain-specific ontologies through four phases. First, we extract domain specific terms from a corpus via machine learning techniques utilizing linguistic patterns (i.e. POS tags). Subsequently, extracted term candidates are filtered through a feature-based approach, and then disambiguated in order to select the most appropriate sense for each term. Third, the filtered and disambiguated terms (in the form of domain concepts) are integrated into existing domain ontologies. In the last phase, a technique is proposed to extract non-taxonomical relations among aforementioned domain concepts. To the best of our knowledge, this is the first

study that incorporates term extraction and relation extraction utilized for ontology enrichment. The feasibility and functionality of the proposed approach is illustrated through a research prototype named *IPO-Extractor* that is built based on a well-accepted NLP platform GATE. *IPO-Extractor* has been demonstrated to successfully deliver all the functionalities of the proposed approach. We evaluate the proposed approach and *IPO-Extractor* in a case study from the finance domain by comparing the results with benchmarks created manually by domain experts. For evaluating the quality of term extraction, we proposed a set of evaluation metrics reflecting the semantics between extracted terms. The results indicate that the proposed method outperforms the extant ontology enrichment mechanisms with respect to both the *quality* (knowledge richness and explicitness) and *efficiency* (computational complexity and resource dependency).

In terms of IPO pricing predictions, we focus on studying the impacts of management's awareness of risks (reflected in the *Risk Factors* sections in the prospectuses) and management's confidence (reflected in the *MD&A* sections in the prospectuses) on IPO valuations. The key contributions of the study include an analytical framework (refer Section 4.1), instantiated in the form a text analytics system (refer Figure 5.1). A text analytic pipeline called *FOCAS-IE* is presented using the analytics system that integrates standard NLP and IE components, while including novel components proposed in this research. One of the novel components is the reasoning component in *FOCAS-IE*, and the associated metrics to measure management's awareness of risks and its outlook of the organization's future performances. These metrics quantify aforementioned hard-to-measure determinants of IPO pricing. The values of these metrics are then utilized in predictive models, which have shown to yield good results. The overall analytics process provides insights in the investment decision making and underwriting processes. Based on the study, several recommendations are relevant for the practitioners:

- Investors and underwriters should spend more time in reading/writing the *Risk Factors* and *MD&A* sections;
- Among the *Risk Factors* and the *MD&A* sections, the informative contents (the *emphasized* mentions of the *risk-features* and the *forward-looking, sentiment-*

subjective mentions of the *MD&A features*), convey more predictive power toward pre- and post-IPO pricing trends;

- In terms of predicting post-IPO pricing trends, the informative contents from the *Risk Factors* sections are more insightful, comparing to the informative contents from the *MD&A* sections;
- On the other hand, MD&A sections are more informative predicting pre-IPO price revisions;
- Pre-IPO price revisions possess substantial effects on the post-IPO initial returns of the same issuance.

Such findings provide guidelines for practitioners such as non-expert investors and under-experienced underwriters, in making investment decisions and writing IPO prospectuses. The results proposed in this study also provide the fundamentals for more advanced analysis, such as understanding the explanatory power of each feature toward the IPO pricing trends.

6.2 Contributions Revisited

In this section, we revisit the contributions from this research study by evaluating the design artifacts against the design requirements mentioned in Section 2.4. Each of the design requirements are discussed below in light of the research conducted.

- With the prospectus ontology created, we have established a formalized conceptualization for knowledge in this specific domain. All the relevant concepts mentioned earlier, along with the relations between them and the properties describing them would be included in the ontology through ontology instantiation and learning. An automated population mechanism has also been developed in order to update the domain knowledge within the ontology.
- With the assistance of GATE (particularly ontology-based gazetteers and APOLDA), various textual representations of relevant entities can be extracted from prospectus document, as well as the extent of mentions.
- With the subsumption reasoning enabled by the ontology, the relations between relevant domain concepts could be identified. Also, the semantic reasoning facilitates extracting other hidden linkages for predicting pre- and post-IPO stock prices.

- With the ontology population and learning mechanism, it is possible to decrease manual effort in the annotation, reasoning, and analytics activities.
- A JAPE rule has been written for dividing the prospectus documents into different sections (e.g. *prospectus summary*, *risk factors*, *MDA*, etc.) in order to enable contextual analysis.
- With the prospectus ontology created, the extracted knowledge is incrementally added to it in the form of ontological classes, relations, or properties. Further, the results from the reasoning and querying activities are also represented in these forms. Thus, the outcome artifacts in formalized representations can be directly imported to the domain knowledge base.
- With the help of predictive modeling techniques, we are able to predict the IPO pricing volatility within and right after the IPO process. Moreover, we have isolated the impacts of features from different sections (i.e. *Risk Factors*, *MD&A*) – which could be used as guidelines for investment and underwriting.

6.3 Lessons Learned

During the design and implementation of the proposed framework, several issues have been identified that need to be addressed in related future studies. These issues might lead to future research opportunities in various domains. They may also be further developed into guidelines for developing future similar applications. Lessons learned from this project could be categorized into two groups: method-oriented and application-oriented.

6.3.1 Method-oriented Issues

Method-oriented issues refer to the issues that are related to the technique we used to develop the proposed framework, which is in particular, Ontology based Information Extraction (OBIE).

First, ontology learning has become increasingly important with respect to OBIE process, due to the domain knowledge and expertise required in constructing and managing ontologies. Average end users would benefit from (semi-) automated ontology learning by reducing these barriers via the development of generic guidelines, the adoption of ontology learning achievements in OBIE applications, and the approaches that increase the accuracy and efficiency of ontology learning in OBIE (Gómez-Pérez & Manzano-Macho, 2005). In the proposed framework, we incorporate ontology learning in the reasoning and learning module. We believe that enabling such functionality in the proposed framework would allow updating the domain knowledge base by adding new knowledge nuggets (in the context of IPO in this study).

Second, ontology integration in the context of OBIE should receive more attention. Current OBIE systems often rely on a single ontology (e.g. only the prospectus ontology has been used in the proposed framework), yet typically a business scenario usually involves multiple ontologies (e.g. An “organization” ontology is used to depict the organizational structure of a firm, and an “activity” ontology is used to reflect the tasks within the firm’s business processes). Additional efforts should be devoted to develop new ontology integration approaches in order to leverage issues such as ontology collaboration, conflicts handling, and so forth (Reeve & Han, 2005; Wimalasuriya & Dou, 2009; Wood, Lydon, Tablan, Maynard, & Cunningham, 2004). A *bridging ontology* approach is proposed for such purposes (Xu, Wang, Lu, Li, & Kang, 2004), but there is much room for additional work on this aspect.

Third, novel, project-specific evaluation metrics need to be developed in OBIE projects. Current metrics are mostly adopted from the Information Retrieval (IR) field (such as the ones used in this project: *recall*, *precision*, and *F-measure*). New measures that are exclusive to the process of OBIE, and are derived from the goals of the project, should be developed in order to provide both cross-sectional and case-wise evaluations to OBIE systems.

With respect to ontology learning/enrichment, compared to the extant literature noted in Section 2.3, our approach is novel from two standpoints: i) applying feature-based WSD while using dictionaries is not common in prior methods; ii) our approach suggests the possibility to align domain specific knowledge base(s) (i.e. *IPO-Ontology*)

with domain independent knowledge base(s) (i.e. WordNet). From the latter standpoint, our approach potentially enables domain-specific, ontology-based reasoning using axioms and semantic relations inherited from domain independent resources. On the other side of the coin, since WordNet cannot fully reflect all semantic relations from real-world scenarios, the marriage with domain-specific knowledge resources would further enrich WordNet by adding new properties/relations.

Last but not least, our approach contributes to the literature of relation extraction, which is a latest development in the ontology learning/enrichment domain. Comparing to prior studies, our approach distinguishes itself in three aspects: firstly, our approach does not depend on a voluminous text corpus, such as the approaches in (Fader et al., 2011; Sánchez et al., 2012; Shen et al., 2012), to extract comparably accurate relations. Secondly, our approach is more capable of picking up semantic dependencies, when comparing to the methods relying on association rules, i.e., co-occurrence relationships (T. Jiang et al., 2007), at a finer-grained grammatical level (clauses versus sentences (Hwang et al., 2011)). Finally, our approach provides domain specific, actionable relations, comparing to similar studies such as (Corro & Gemulla, 2013; Punuru & Chen, 2011), which better supports further reasoning and other knowledge-intensive analyses.

Other than aforementioned issues, attention should be given to developing intelligent user interfaces to lower the adoption barriers, enhancing change control toward ontologies in OBIE projects, improving the performance of co-reference resolution through the use of ontologies, and so forth.

6.3.2 Application-oriented Issues

The application-oriented issues fall into two main categories: the issues regarding applying the proposed framework in the context of IPO, and the issues regarding adopting similar OBIE applications to other domains.

As discussed in Section 2, the major use cases of the proposed framework include: estimating the short-term price offerings in the IPO process, and providing referential support for the underwriters when composing an IPO prospectus. In regards to the aspect of price estimation, the proposed framework would enhance the studies aimed at understanding the pricing phenomena (such as the “*underpricing*” phenomena and the

anomalies in price offerings, etc.) by providing: 1) a formalized conceptualization that is more reliable and efficient when comparing to the manual efforts of the domain experts; 2) a mechanism that enables the automated processing of the prospectus, which is highly suggested as one of the most significant future steps in Bhabra & Pettway (2003) and more thorough knowledge discovery within the prospectus with consideration of cross-version and cross-section analyses – comparing to the SVM-based method proposed in Hanley & Hoberg (2010).

Additionally, the prospectus ontology would become a rich domain knowledge base via the ontology learning and ontology instantiation features from the proposed framework. A set of practical guidelines derived from the prospectus ontology aimed at providing decision support to the underwriters assessing prospectus documents needs to be formalized.

The proposed techniques in this work also allow further analyses toward the IPO pricing phenomenon through mining IPO prospectuses. One possible analysis of this kind is “delta analysis”. *Delta analysis* refers to tracking changes or identifying the “gaps” between different versions. During the time span from the release of the first version of the prospectus to the final offering, several versions of the prospectus are issued to the public. Based on the premise that the content changes are meaningful for the knowledge discovery purpose, these “gaps” should be analyzed. Through a preliminary study, a freeware tool named *daisydiff* has been employed for comparing different versions of the prospectus; then a JAPE rule is coded to annotate what have been changed and how they have been changed (e.g. *added/removed*, etc.) in the contents. Such analysis would provide additional insights to the IPO pricing phenomenon. A brief demonstration of the delta analysis can be found in Figure 6.1 below.

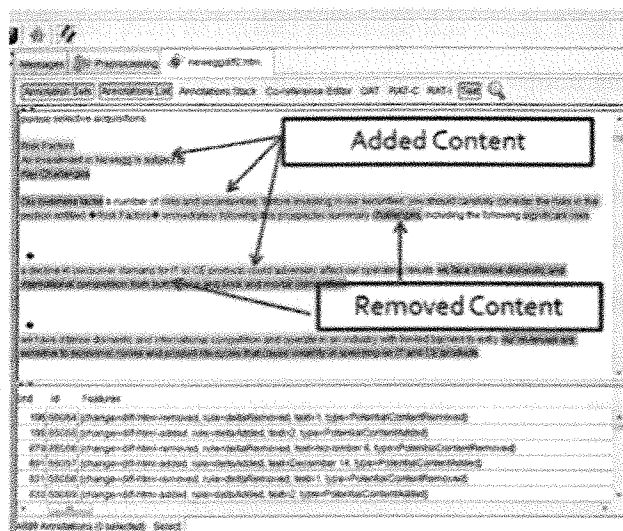


Figure 6.1. “Delta Analysis” on IPO Prospectuses

At a more global level, applications similar to the proposed framework could be used in a variety of domains, including bioinformatics, medicine, semantic web, and process analytics (i.e. process mining, policy enforcement, business intelligence).

6.4 Limitations and Future Steps

As with any research study, there are limitations and shortcomings of this study as well, which need to be acknowledged. Consider the ontology enrichment research component. First, the relatively low quality performance for relation extraction could partially be explained by the fact that there are only a few well-accepted, machine-readable concepts and axioms in the finance domain. For instance, in the widely-used US GAAP, a term “revenue” would have more than 20 variations, while only less than 10 axioms are defined between it and other related concepts (Chowdhuri, Yoon, Redmond, & Etudo, 2014). This variation is a challenge since we do not have a sufficient foundation for enrichment tasks, but it also indicates a good opportunity – which is part of the reason we selected this domain in our case study. Second, the proposed WSD method is not exhaustive in the sense that some of the terms could not be disambiguated. As a subsequent step of this study, we will experiment enhancing the proposed WSD method with relations from existing domain knowledge bases as well as extracted from the text

corpus. In addition, we also plan to further enhance this study from an application perspective. This study is a component of a larger project, which aims at studying the IPO pricing strategies based on the IPO prospectus. The approach proposed in this paper prepares the basis for further analysis. In the future, we plan to: apply the approach or its improved variants to textual contents in other important sections in the IPO prospectus, and then construct the ontology for the overall IPO process; and also use the enriched ontology as the basis to extract knowledge from the IPO prospectus, and then use such knowledge for constructing predictive models for understanding the IPO pricing phenomenon.

Consider the predictive modeling research component. We recognize the need to improve our reasoning mechanism to better depict the sentiments at sentence level by reducing the overlapping across features/sentiments, so that we can provide an even more accurate proxy of the management's tone. This could be achieved using advanced text analytics techniques to improve the design artifacts proposed in this study. We are considering using artificial intelligence (AI) based document-understanding techniques for this purpose. From a validation standpoint, we need to further test and apply our approach on an even larger and more comprehensive dataset. The sample size need to be increased, and with diverse IPO prospectus variants. We are working on extending the study to include other key sections of the IPO prospectuses. Another future direction under consideration is to investigate other types of IPO prospectus (Form S-1 or other Form 424 variants) documents. Also, as suggested in Section 5.4.4, discovering more advanced feature aggregation (dimension reduction) approaches might mitigate the decline in prediction power.

REFERENCES

- Aizawa, A. (2003). An information-theoretic perspective of tf-idf measures. *Information Processing & Management*, 39(1), 45–65. doi:10.1016/S0306-4573(02)00021-3
- Apache-OpenNLP-Development-Community. (2014). Apache OpenNLP Developer Documentation. Retrieved from <https://opennlp.apache.org/documentation/1.5.3/manual/opennlp.html>
- Appelt, D. (1999). An introduction to information extraction. *Artificial Intelligence Communications*, 12(3), 161–172.
- Arnold, T., Fishe, R. P. H., & North, D. (2010). The Effects of Ambiguous Information on Initial and Subsequent IPO Returns. *Financial Management (Blackwell Publishing Limited)*, 39(4), 1497–1519.
- Asquith, D., Jones, J. D., & Kieschnick, R. (1998). Evidence on Price Stabilization and Underpricing in Early IPO Returns. *Journal of Finance*, 53(5), 1759–1773.
- Babich, V., & Sobel, M. J. (2004). Pre-IPO Operational and Financial Decisions. *Management Science*, 50(7), 935–948. doi:10.1287/mnsc.1040.0252
- Bhabra, H. S., & Pettway, R. H. (2003). IPO Prospectus Information and Subsequent Performance. *The Financial Review*, 38(3), 369–397. doi:10.1111/1540-6288.00051
- Buitelaar, P., Cimiano, P., & Magnini, B. (2005). Ontology Learning from Text: An Overview. In P. Buitelaar, P. Cimiano, & B. Magnini (Eds.), *Ontology Learning from Text Methods Applications and Evaluation* (pp. 3–12). IOS Press.
- Campbell, J. L., Chen, H., Dhaliwal, D. S., Lu, H., & Steele, L. B. (2014). The Information Content of Mandatory Risk Factor Disclosures in Corporate Filings. *Review of Accounting Studies*, 19(1), 396–455.
- Carter, R. B., Dark, F. H., & Singh, A. K. (1998). Underwriter Reputation , Initial Returns , and the Long-Run Performance of IPO Stocks. *The Journal of Finance*, LIII(1), 285–312.
- Chan, S. W. K., & Franklin, J. (2011). A text-based decision support system for financial sequence prediction. *Decision Support Systems*, 52(1), 189–198. doi:10.1016/j.dss.2011.07.003
- Chemmanur, T. J. (1993). The Pricing of Initial Public Offerings : A Dynamic Model with Information Production. *Journal of Finance*, 48(1), 285–305.

- Chowdhuri, R., Yoon, V. Y., Redmond, R. T., & Etudo, U. O. (2014). Ontology based Integration of XBRL Filings for Financial Decision Making. *Decision Support Systems*. doi:10.1016/j.dss.2014.09.004
- Clarkson, P. M., & Thompson, R. E. X. (1990). Empirical Estimates of Beta When Investors Face Estimation Risk. *Journal of Finance*, *XLV*(2), 431–454.
- Cornelli, F., Goldreich, D., & Ljungqvist, A. (2006). Investor Sentiment and Pre-IPO Markets. *Journal of Finance*, *61*(3), 1187–1217.
- Corro, L. Del, & Gemulla, R. (2013). ClausIE : Clause-Based Open Information Extraction. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 355–365).
- Cowie, J., & Wilks, Y. (2000). Information extraction. In R. Dale, H. Moisl, & H. Somers (Eds.), *Handbook of Natural Language Processing* (pp. 241–260). CRC Press.
- Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. (2002). GATE: an architecture for development of robust HLT applications. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 168–175).
- Cunningham, H., Maynard, D., & Tablan, V. (2000). *JAPE: a Java Annotation Patterns Engine*.
- Daily, C. M., Certo, S. T., Dalton, D. R., & Roengpitya, R. (2003). IPO Underpricing: A Meta-Analysis and Research Synthesis. *Entrepreneurship: Theory & Practice*, *27*(3), 271–296.
- Dorji, T. C., Atlam, E., Yata, S., Fuketa, M., Morita, K., & Aoe, J. (2010). Extraction, selection and ranking of Field Association (FA) Terms from domain-specific corpora for building a comprehensive FA terms dictionary. *Knowledge and Information Systems*, *27*(1), 141–161. doi:10.1007/s10115-010-0296-x
- Fader, A., Soderland, S., & Etzioni, O. (2011). Identifying Relations for Open Information Extraction. In *Proceedings of EMNLP '11 Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1535–1545).
- Ferris, S. P., Hao, G. Q., & Liao, S. M. (2012). The Effect of Issuer Conservatism on IPO Pricing and Performance. *Review of Finance*, 1–45.
- Ferris, S. P. S., Hao, Q., & Liao, M.-Y. (2013). The Effect of Issuer Conservatism on IPO Pricing and Performance*. *Review of Finance*, *17*(3), 993–1027. doi:10.1093/rof/rfs018

- Firth, M., Wang, K., & Sonia, W. (2013). Corporate Transparency and the Impact of Investor Sentiment on Stock Prices. *Management Science*, *Forthcoming*, (September).
- Gaeta, M., Orciuoli, F., Paolozzi, S., & Salerno, S. (2011). Ontology Extraction for Knowledge Reuse: The e-Learning Perspective. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *41*(4), 798–809. doi:10.1109/TSMCA.2011.2132713
- Ge, J., & Qiu, Y. (2008). Concept Similarity Matching Based on Semantic Distance. In *2008 Fourth International Conference on Semantics, Knowledge and Grid* (pp. 380–383). IEEE. doi:10.1109/SKG.2008.24
- Geva, T., & Zahavi, J. (2014). Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news. *Decision Support Systems*, *57*, 212–223. doi:10.1016/j.dss.2013.09.013
- Gómez-Pérez, A., & Manzano-Macho, D. (2005). An overview of methods and tools for ontology learning from texts. *The Knowledge Engineering Review*, *19*(03), 187–212. doi:10.1017/S0269888905000251
- Gooch, P. (2013). GATE plugin for adding WordNet features to annotations. Retrieved from https://github.com/philgooch/WordNet_Suggester
- Grishman, R. (1997). Information extraction: Techniques and challenges. *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology Lecture Notes in Computer Science*, *1299*, 10–27.
- Hagenau, M., Liebmann, M., & Neumann, D. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems*, *55*(3), 685–697. doi:10.1016/j.dss.2013.02.006
- Hanley, K. W., & Hoberg, G. (2010). The Information Content of IPO Prospectuses. *Review of Financial Studies*, *23*(7), 2821–2864.
- Hanley, K. W., & Hoberg, G. (2012). Litigation Risk , Strategic Disclosure and the Underpricing of Initial Public Offerings. *Journal of Financial Economics*, *103*(2), 235–254.
- Hevner, A. R. A., March, S. T. S., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75–105.
- Hwang, M., Choi, C., & Kim, P. (2011). Automatic Enrichment of Semantic Relation Network and Its Application to Word Sense Disambiguation. *IEEE Transactions on Knowledge and Data Engineering*, *23*(6), 845–858.

- Ittoo, A., & Bouma, G. (2013a). Minimally-supervised learning of domain-specific causal relations using an open-domain corpus as knowledge base. *Data & Knowledge Engineering*, 88, 142–163. doi:10.1016/j.datak.2013.08.004
- Ittoo, A., & Bouma, G. (2013b). Term extraction from sparse, ungrammatical domain-specific documents. *Expert Systems with Applications*, 40(7), 2530–2540. doi:10.1016/j.eswa.2012.10.067
- Jain, B. a., & Kini, O. (1999). The Life Cycle of Initial Public Offering Firms. *Journal of Business Finance and Accounting*, 26(9-10), 1281–1307. doi:10.1111/1468-5957.00298
- Jenkinson, T., & Jones, H. (2009). IPO pricing and allocation: a survey of the views of institutional investors. *Review of Financial Studies*, 22(4), 1477–1504.
- Jiang, T., Tan, A., & Wang, K. (2007). Mining Generalized Associations of Semantic Relations from Textual Web Content. *IEEE Transactions on Knowledge and Data Engineering*, 19(2), 164–179.
- Jiang, X., & Tan, A. (2005). Mining Ontological Knowledge from Domain-Specific Text Documents. *Fifth IEEE International Conference on Data Mining (ICDM'05)*, 665–668. doi:10.1109/ICDM.2005.97
- Jiang, X., & Tan, A. (2006). Learning and inferencing in user ontology for personalized semantic web services. *Proceedings of the 15th International Conference on World Wide Web - WWW '06*, 1067. doi:10.1145/1135777.1136018
- Kang, N., van Mulligen, E. M., & Kors, J. a. (2011). Comparing and combining chunkers of biomedical text. *Journal of Biomedical Informatics*, 44(2), 354–60. doi:10.1016/j.jbi.2010.10.005
- Kang, Y.-B., Delir Haghighi, P., & Burstein, F. (2014). CFinder: An intelligent key concept finder from text for ontology development. *Expert Systems with Applications*, 41(9), 4494–4504. doi:10.1016/j.eswa.2014.01.006
- Khansa, L., & Liginlal, D. (2011). Predicting stock market returns from malicious attacks: A comparative analysis of vector autoregression and time-delayed neural networks. *Decision Support Systems*, 51(4), 745–759. doi:10.1016/j.dss.2011.01.010
- Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. In *WordNet: An electronic lexical database* (pp. 265–283). MIT press.
- Lewellen, K. (2006). Risk , Reputation , and IPO Price Support. *Journal of Finance*, 61(2), 613–654.

- Li, F. (2010). The Information Content of Forward-Looking Statements in Corporate Filings-A Naïve Bayesian Machine Learning Approach. *Journal of Accounting Research*, 48(5), 1049–1102. doi:10.1111/j.1475-679X.2010.00382.x
- Li, J., Wang, H., & Khan, S. U. (2013). A Fully Distributed Scheme for Discovery of Semantic Relationships. *IEEE TRANSACTIONS ON SERVICES COMPUTING*, 6(4), 457–469.
- Li, Q., & Wu, Y.-F. B. (2006). Identifying important concepts from medical documents. *Journal of Biomedical Informatics*, 39(6), 668–79. doi:10.1016/j.jbi.2006.02.001
- Lin, D. (1998). An information-theoretic definition of similarity. In *Proceeding ICML '98 Proceedings of the Fifteenth International Conference on Machine Learning* (pp. 296 – 304).
- Lin, M.-C., Lee, A. J. T., Kao, R.-T., & Chen, K.-T. (2011). Stock price movement prediction using representative prototypes of financial reports. *ACM Transactions on Management Information Systems*, 2(3), 1–18. doi:10.1145/2019618.2019625
- Ljungqvist, A., Nanda, V., & Singh, R. (2006). Hot Markets, Investor Sentiment, and IPO Pricing*. *The Journal of Business*, 79(4), 1667–1702.
- Loughran, T., & McDonald, B. (2011). When is a Liability not a Liability ? Textual Analysis , Dictionaries , and 10-Ks. *Journal of Finance*, 66(1), 35–65.
- Loughran, T., & McDonald, B. (2013). IPO First-Day Returns, Offer Price Revisions, Volatility, and Form S-1 Language. *Journal of Financial Economics*, 109(2), 307–326. doi:10.1016/j.jfineco.2013.02.017
- Loughran, T., & Ritter, J. R. (2004). Why Has IPO Underpricing Increased Over Time? *Financial Management*, 33(3), 1–47.
- Lowry, M. (2003). Why does IPO volume fluctuate so much ? *Journal of Financial Economics*, 67(1), 3–40.
- Lowry, M., & Schwert, G. W. (2004). Is the IPO pricing process efficient? *Journal of Financial Economics*, 71(1), 3–26. doi:10.1016/S0304-405X(03)00205-8
- Maslennikov, M., & Chua, T.-S. (2010). Combining relations for information extraction from free text. *ACM Transactions on Information Systems*, 28(3), 1–35. doi:10.1145/1777432.1777437
- Meijer, K., Frasincar, F., & Hogenboom, F. (2014). A semantic approach for extracting domain taxonomies from text. *Decision Support Systems*, 62, 78–93. doi:10.1016/j.dss.2014.03.006

- Nassif, H., Woods, R., Burnside, E., Ayvaci, M., Shavlik, J., & Page, D. (2009). Information Extraction for Clinical Data Mining: A Mammography Case Study. In *2009 IEEE International Conference on Data Mining Workshops* (pp. 37–42). Ieee. doi:10.1109/ICDMW.2009.63
- Navigli, R. (2009). Word sense disambiguation: A Survey. *ACM Computing Surveys*, 41(2), 1–69. doi:10.1145/1459352.1459355
- Ngo-Ye, T. L., & Sinha, A. P. (2014). The influence of reviewer engagement characteristics on online review helpfulness: A text regression model. *Decision Support Systems*, 61, 47–58. doi:10.1016/j.dss.2014.01.011
- Princeton-University. (2012). About WordNet. Retrieved from <http://wordnet.princeton.edu/>
- Punuru, J., & Chen, J. (2011). Learning non-taxonomical semantic relations from domain texts. *Journal of Intelligent Information Systems*, 38(1), 191–207. doi:10.1007/s10844-011-0149-4
- Rajan, R., & Servaes, H. (1997). Analyst following of initial public offerings. *The Journal of Finance*, 52(2), 507–529.
- Reeve, L., & Han, H. (2005). Survey of semantic annotation platforms. In *Proceedings of the 2005 ACM symposium on Applied computing - SAC '05* (pp. 1634–1638). New York, New York, USA: ACM Press. doi:10.1145/1066677.1067049
- Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (Vol. 1).
- Ritter, J. R. (1991). The Long-Run Performance of Initial Public Offerings. *The Journal of Finance*, 46(1), 3–27.
- Ritter, J. R., & Welch, I. (2002). A Review of IPO Activity, Pricing, and Allocations. *Journal of Finance*, LVII(4), 1795–1828.
- Roosenboom, P., & Thomas, J. (2007). How Do Underwriters Value Initial Public Offerings? An Empirical Analysis of the French IPO Market. *Contemporary Accounting Research*, 24(4), 1217–1243. doi:10.1506/car.24.4.7
- Ruiz-Martínez, J. M., Valencia-García, R., Martínez-Béjar, R., & Hoffmann, A. (2012). BioOntoVerb: A top level ontology based framework to populate biomedical ontologies from texts. *Knowledge-Based Systems*, 36, 68–80. doi:10.1016/j.knosys.2012.06.002

- Sánchez, D., Batet, M., Valls, A., & Gibert, K. (2009). Ontology-driven web-based semantic similarity. *Journal of Intelligent Information Systems*, 35(3), 383–413. doi:10.1007/s10844-009-0103-x
- Sánchez, D., Moreno, A., & Del Vasto-Terrientes, L. (2012). Learning relation axioms from text: An automatic Web-based approach. *Expert Systems with Applications*, 39(5), 5792–5805. doi:10.1016/j.eswa.2011.11.088
- Schmitz, M., Bart, R., Soderland, S., & Etzioni, O. (2012). Open Language Learning for Information Extraction. In *Proceeding of EMNLP-CoNLL '12 Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 523–534). ACM.
- Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news. *ACM Transactions on Information Systems*, 27(2), 1–19. doi:10.1145/1462198.1462204
- Sen, A., Dacin, P. A., & Pattichis, C. (2006). Current Trends in Web Data Analysis. *Communications of the ACM*, 49(11), 85–91.
- Sen, S., Tao, J., & Deokar, A. V. (2014). On the Role of Ontologies in Information Extraction. *Annals of Information Systems*.
- Shen, M., Liu, D.-R., & Huang, Y.-S. (2012). Extracting semantic relations to enrich domain ontologies. *Journal of Intelligent Information Systems*, 39(3), 749–761. doi:10.1007/s10844-012-0210-y
- Shima, H. (2013). WS4J. Retrieved from <https://code.google.com/p/ws4j/>
- Snowball-Tartarus. (2013). English stop word list. Retrieved from <http://snowball.tartarus.org/algorithms/english/stop.txt>
- Song, M., Yang, H., Siadat, S. H., & Pechenizkiy, M. (2013). A comparative study of dimensionality reduction techniques to enhance trace clustering performances. *Expert Systems with Applications*, 40(9), 3722–3737. doi:10.1016/j.eswa.2012.12.078
- Tang, H., Liao, S. S., & Sun, S. X. (2013). A prediction framework based on contextual data to support Mobile Personalized Marketing. *Decision Support Systems*, 56, 234–246. doi:10.1016/j.dss.2013.06.004
- Varshney, S., & Robinson, R. (2004). IPO Research Symposium Review. *Journal of Economics and Finance*, 28(1), 56–67.
- Wartena, C., Brussee, R., Gazendam, L., & Huijsen, W.-O. (2007). Apolda: A practical tool for semantic annotation. In R. Brussee (Ed.), *Proceedings of the 18th*

International Conference on Database and Expert Systems Applications (DEXA '07) (pp. 288–292).

- Wasikowski, M., & Chen, X. (2010). Combating the small sample class imbalance problem using feature selection. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1388–1400.
- Weichselbraun, A., Wohlgenannt, G., & Scharl, A. (2010). Refining non-taxonomic relation labels with external structured data to support ontology learning. *Data & Knowledge Engineering*, 69(8), 763–778. doi:10.1016/j.datak.2010.02.010
- Wimalasuriya, D. C., & Dou, D. (2009). Using multiple ontologies in information extraction. In *Proceeding of the 18th ACM conference on Information and knowledge management - CIKM '09* (pp. 235–244). New York, New York, USA: ACM Press. doi:10.1145/1645953.1645985
- Wimalasuriya, D. C., & Dou, D. (2010). Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science*, 36(3), 306–323. doi:10.1177/0165551506nnnnnn
- Wimmer, H., & Zhou, L. (2013). Word Sense Disambiguation for Ontology Learning. In *Proceedings of the Nineteenth Americas Conference on Information Systems* (pp. 1–10).
- Wong, W., Liu, W., & Bennamoun, M. (2009). Acquiring Semantic Relations Using the Web for Constructing Lightweight Ontologies, 266–277.
- Wong, W., Liu, W., & Bennamoun, M. (2012). Ontology learning from text: A look back and into the future. *ACM Computing Surveys*, 44(4), 1–36. doi:10.1145/2333112.2333115
- Wood, M. M., Lydon, S. J., Tablan, V., Maynard, D., & Cunningham, H. (2004). Populating a Database from Parallel Texts Using Ontology-Based Information Extraction. In *Natural Language Processing and Information Systems* (pp. 254–264).
- Wu, Z., & Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics (ACL '94)* (pp. 133–138).
- Xu, B., Wang, P., Lu, J., Li, Y., & Kang, D. (2004). Bridge Ontology and Its Role in Semantic Annotation. In *Proceedings of the International Conference on Cyberworlds (CW '04)* (pp. 329–334).

- Zhang, C., Niu, Z., Jiang, P., & Fu, H. (2012). Domain-specific term extraction from free texts. *2012 9th International Conference on Fuzzy Systems and Knowledge Discovery*, (Fskd), 1290–1293. doi:10.1109/FSKD.2012.6234350
- Zhang, Z. (2008). Mining relational data from text: From strictly supervised to weakly supervised learning. *Information Systems*, 33(3), 300–314.
- Zhong, N., Li, Y., & Wu, S. (2012). Effective Pattern Discovery for Text Mining. *IEEE Transactions on Knowledge and Data Engineering*, 24(1), 30–44.
- Zhou, L. (2007). Ontology learning: state of the art and open issues. *Information Technology and Management*, 8(3), 241–252. doi:10.1007/s10799-007-0019-5

Appendix A. Research Plan: Data

Analytics Lifecycle

From the point of view of project management and execution, the research study follows the data analytics lifecycle, which is described in here.

Given that the two key areas in this project, namely text analytics and predictive modeling, fall under the overall realm of data analytics, we found it logical to adopt data analytics life cycle (DALC) to ensure the quality of the outcomes. DALC refers to the process, or the “de facto” procedure, guiding any data analytics project. A variety of DALC models are employed in the field; among them, two DALC models are accepted more widely than the others, namely Cross Industry Standard Process for Data Mining (CRISP-DM) (c.f.

http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining) and EMC’s Data Analytics Lifecycle (EMC-DALC) (c.f.

https://infocus.emc.com/david_dietrich/the-genesis-of-emcs-data-analytics-lifecycle/).

Based on these two models, I adapted the DALC for this project as shown in Figure A.1.

The DALC depicted in Figure A.1 was followed closely in this project as the research plan. First, a pilot study was conducted in order to shed light on the design/analytical problem (IPO Pricing) – manual and semi-automatic coding and analysis are used in this step, which helped formulate the design/analytical problem better.

Next, we collected data from data sources, and then prepared it for respective data analysis purposes. Data collection in this project was conducted in a two-phase manner. A web crawler was developed to incrementally retrieve IPO prospectuses from SEC EDGAR database, which comprises the text corpus used in this study. The retrieved data were filtered and cleaned. Then, based on different analytical purposes, different text analytics/predictive analytics techniques were used for data cleaning and preparation. For instance, semantic annotation is used for ontology learning purposes, while IE is used for

reasoning purposes (which provides data for predictive modeling). Details regarding these steps can be found in Section 4.3 and 4.4.

In the third step, the actual analytical models were planned and constructed. Analytical models in this project majorly refer to predictive models, which were used to explain/predict pre- and post-IPO pricing trends. Tasks conducted in this step include (but not limit to): representation of the analytical problem, identification of different types of variables (i.e. *predictor*, *target*, *control*), and selection and prototyping of different predictive modeling techniques.

Three major steps were conducted in the next step, namely *Predictive Modeling*, *Model Evaluation*, and *Results Communication*. The constructed predictive models from previous step were executed with data extracted and quantified from the analytical framework. Different portions of the data, as well as the whole dataset, were used in a series of experiments. The experiment results from the predictive models were evaluated via two dimensions: *accuracy* and *performance*, which are widely adopted in evaluating predictive results. Then the evaluated results were interpreted in the context of IPO pricing prediction, to ensure communication to practitioners and academic researchers in the finance and related domains. Details regarding this step can be found in Chapter 5.

Last but not least, the predictive models, with associated approaches, need to be deployed and operationalized. As discussed above, the approaches and the predictive models have been realized in research prototypes.

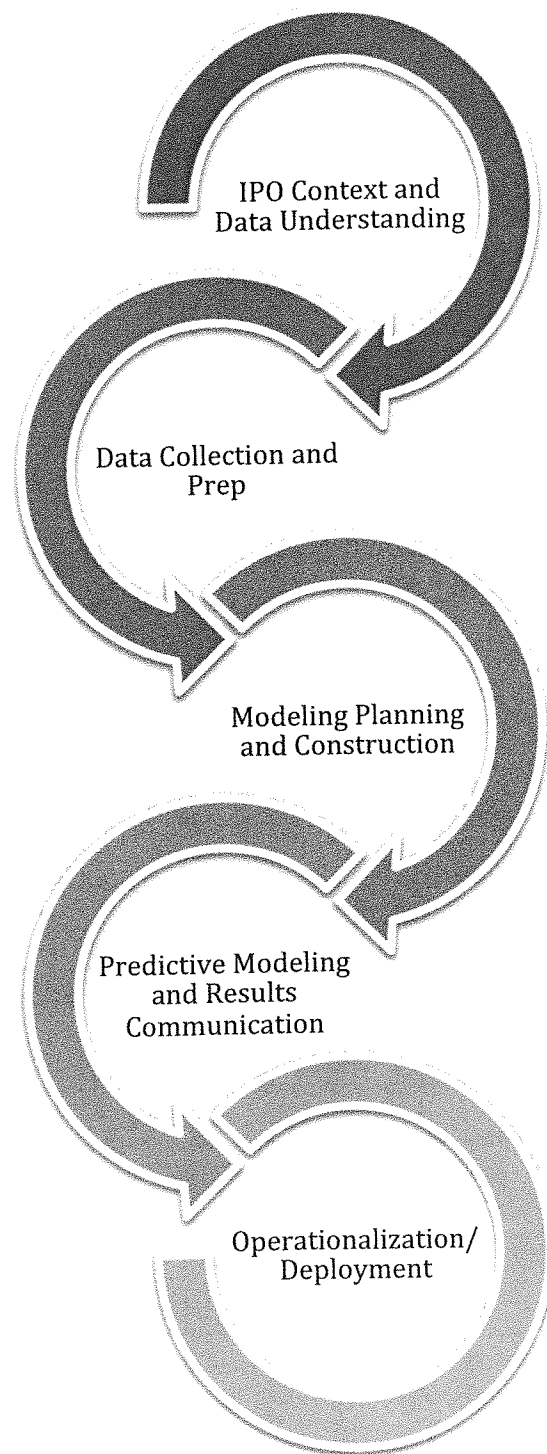


Figure A.1 DALC used in This Project as Research Plan