3/10

2021 - Master Thesis - 58 pages - 2021.11.01

- very "watery", amateurish → MBA non-finance

- non-programmer, non computer science at all...

- relevant research topic

- overall not impressed

IOWA STATE UNIVERSITY
Digital Repository

Creative Components

Iowa State University Capstones, Theses and Dissertations

Spring 2021

Leveraging Text Mining and Analytical Technology to Enhance Financial Planning and Analysis

Yih-Shan Sheu

Follow this and additional works at: https://lib.dr.iastate.edu/creativecomponents

 Part of the Business Analytics Commons

Recommended Citation

Sheu, Yih-Shan, "Leveraging Text Mining and Analytical Technology to Enhance Financial Planning and Analysis" (2021). *Creative Components*. 810.
https://lib.dr.iastate.edu/creativecomponents/810

This Creative Component is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Creative Components by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.

Leveraging Text Mining and Analytical Technology to Enhance Financial Planning and Analysis

by

Yih-Shan Sheu

A creative component submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

MASTER OF BUSINESS ADMINISTRATION (MBA)

and

MASTER OF SCIENCE (MS)

Major: Business Analytics (MBA) and Information Systems (MS)

Program of Study Committee:
Dr. Anthony M. Townsend, Major Professor (MSIS)
Dr. Valentina Salotti (MBA)

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this creative component. The Graduate College will ensure this creative component is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2021

Copyright © Yih-Shan Sheu, 2021. All rights reserved.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

Big data technologies have substantially affected various industries. Though data science has been the most valuable evolution in the age of technological innovation, the financial sector is lagging behind other sectors through leveraging data science to evolve quickly and emphasize competency in data analytics. Although big data technology used in financial services, such as FinTech and stock trending models, has grown immensely in the past few years, there is still little research in Corporate Finance. This paper focuses on the big-data technology application in corporate finance via ***text mining*** and ***algorithmic forecasting model***. This study aims to answer the following two research questions: (i) How to handle unstructured information to gain an in-depth understanding of qualitative data that will impact the financial performance; (ii) How could machine learning help Corporate Finance acquire better market trend insights and achieve precise sales prediction as well as financial forecasting? In order to answer these questions, a qualitative analysis of literature is carried out comprehensively. Recent research and study indicate that such applications in corporate finance can significantly benefit the corporate decision-making process due to more timely, more relevant, and customer-oriented factors involving qualitative data sources. Finally, the paper briefly discusses the current challenges and limitations and points out the potential future scope of data technology in corporate finance.

1. INTRODUCTION

Today, finance talent with a robust professional background is no longer sufficiently competitive to become a value-added business partner over the long term.

Data science[1] has been emerging within business sectors such as financial markets, internet banking, and financial service management. "Data science" in this paper would refer to specific

[1]Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data. Data science is related to data mining, machine learning and big data (Wikipedia).

techniques that use machine learning, data mining, or big data to extract knowledge and insights from many structural and unstructured data. There is a growing tendency that more and more financial operations leverage automation with predefined rules, such as risk assessment, algorithmic execution, or pricing model, to make decisions. They transform the financial markets into hyperconnected networks for ultra-fast exchanging information (Neevista Pty Ltd, 2020). We learn that data science techniques have played an essential role in improving the financial services sector, particularly in trade and investment, risk analysis, fraud detection and investigation, tax reform, and automation (Hasan et al., 2020). However, this technique is still an extremely new concept within corporate finance departments. Nowadays, machine learning has become more accessible, thanks to various no-code, low-code tools on the market. Some well-built platforms, such as RapidMiner, allow companies to quickly perform data mining using pre-trained models or creating customized solutions in a user-friendly interface. Finance talent could develop technology acumen via such tools to create more value in their planning analysis.

Business strategic planning relies heavily on accurate forecasting, which helps market trend prediction and creates a more explicit focus for future objectives. Traditional business forecasting relies on historical quantitative data and manual spreadsheets, prone to human bias and errors. Financial analysis, in particular, requires more precise data for better strategic planning; therefore, financial analysts in corporate finance should use external and alternative data such as industry data and microeconomic information to do better planning. Nevertheless, data science such as Artificial Intelligence (AI) and other big data techniques have not yet gained traction within corporate finance functions and have not created much headway in ensuring the value-added nature of the big datasets and data techniques. The goals of business forecasting are to provide benchmarks and to reduce uncertainty for monitoring actual performance. Emerging information technologies will improve prediction accuracy and contribute to enhancing the bottom line.

This research focuses on leveraging data science with finance activities in corporate finance

(budgeting/ forecasting/ reporting/ strategic planning / creating dashboards, etc.) by concentrating on text mining application and algorithmic financial forecasting from different dimensions. Corporate finance is an essential aspect of each company's financial domain because it integrates its overall functioning with its financial structure. Data science can be incorporated into financial modeling to create better predictive analysis and more in-depth insights through identifying more accurate and robust assumptions. The development of Enterprise Resource Planning (ERP) systems is a significant step in that evolutionary process. Before ERP deployment, it is incredibly inefficient to analyze various data across massive separate isolated databases. ERP systems have boosted incremental optimizations within business operations, including finance functions, in the last decade. For instance, Shared Services Center (SSC) is a crucial business process reengineering in the past ten years. With the shared increasing data collection and data mining, finance departments surely could benefit from data technology and improve their business partnering capabilities.

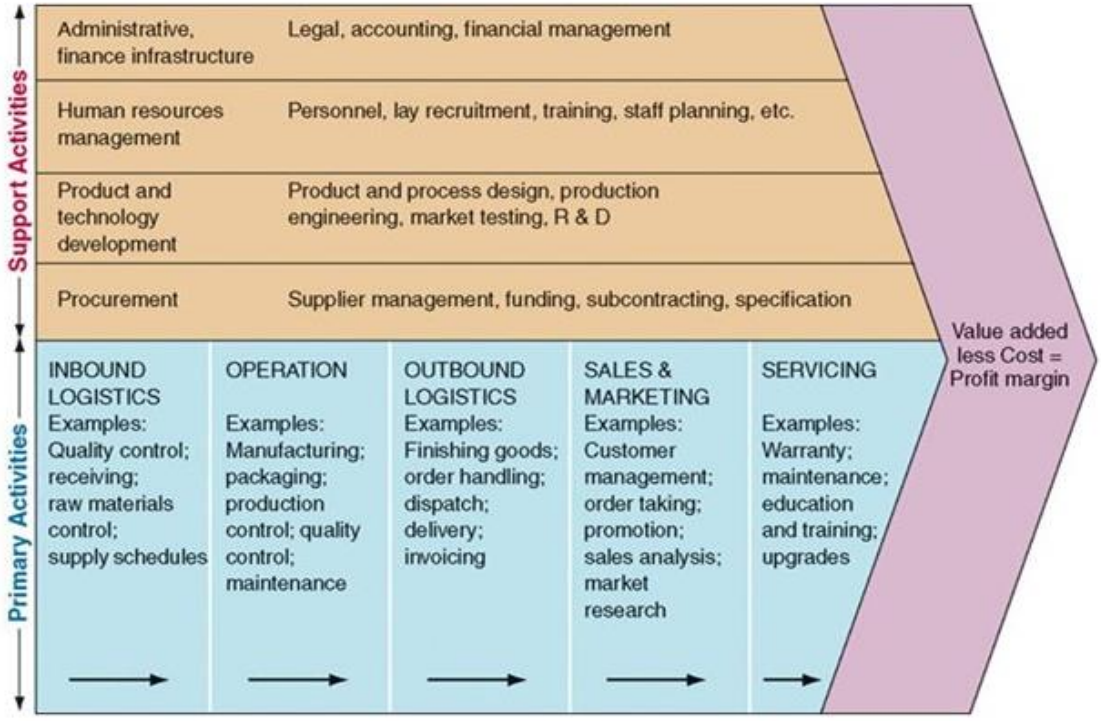
2. RESEARCH QUESTIONS (RQs)

Corporate finance collects and generates a vast amount of data from different dimensions in a company. From Porter's Value Chain's perspectives[2], corporate finance analyzes data on each value chain like procurement process, production efficiency, supply chain management (delivering the products or service), marketing and channel strategy, and sales performance [Figure 1]. Currently, structured data is easy to store in a finance Hyperion system. Semi-structured or unstructured data from various corporate documents such as the annual reports (10-K) or Management discussion and analysis (MD&A) of a company have a lot of hidden economic context information and are under-evaluated. Today, text analytics is pervasively used in

[2] Porter's Value Chain Analysis is a business management concept that was developed by Michael Porter. In his book Competitive Advantage (1985), Michael Porter explains Value Chain Analysis; that a value chain is a collection of activities that are performed by a company to create value for its customers.

customer relationship management (CRM) to measure customer opinions and satisfaction. Some researchers, such as Gupta et al. (2020), believe text-mining techniques can be employed to extract this hidden financial information from annual reports and predict future financial sustainability in more precise mechanisms. According to the June 2020 The CPA Journal, textual analysis for corporate 10-K filings by applying a contemporary Natural Language Processing (NLP) approach could empower users to understand textual documents in different dimensions and create corporate risk profiles. As an essential branch of artificial intelligence, a natural language processing (NLP) approach enables computers to understand the human language on a large scale to analyze unstructured textual documents in financial reporting and predict future economic sustainability (Liu et al., 2020).

Figure 1- Porter's Value Chain



Source:https://www.mbaskool.com/business-concepts/marketing-and-strategy-terms/2516-porter-value-chain.html

This paper is going to shed more light on the following research questions through an extensive literature review. As noted earlier, data science techniques are underutilized in finance

processes and could create better business outcomes. Therefore, this paper hypothesizes that:

- *Hypothesis 1: Using data science to handle semi-structured or unstructured information will create better business outcomes.*

Therefore, the first question is:

- RQ1: How could data science handle unstructured information to gain an in-depth understanding of qualitative data that will potentially impact the financial performance and forecasting model? In other words, how is the textual analysis for the financial reports trying to quantify textual data as meaningful predictors for future financial performance?

"Data science" is a vast and somewhat nebulous term, and this research is particularly interested in machine learning techniques. It is hypothesized that:

- *Hypothesis 2: Machine learning for finance is applicable and accessible, and beneficial to corporate finance in key areas such as strategic planning, budgeting, and forecasting.*

Therefore, the second research question is:

- RQ2: How could machine learning help Corporate Finance acquire better market trend insights and achieve precise sales prediction as well as financial forecasting?

In order to provide the answers to these two research questions (RQs) for this study, several key pieces of information need to be taken into consideration when conducting comprehensive financial forecasting. That includes quantitative and qualitative information from both internal and external the individual company and the larger industry the company is situated in. First, traditionally unstructured information is facing challenges in being disclosed or evaluated. Qualitative information is undermined in the forecasting process. Big data sentiment analysis has potential applications in sales forecasting (Lau et al., 2018). Negative sentiment analysis appearing in online news, social media, and other online sources may aim to uncover people's attitudes (e.g., positive or negative) toward entities or products. For example, consistent negative

sentiment about particular products might steer financial planners to reevaluate whether sales forecasting is too optimistic (Gepp et al., 2018). Secondly, top-line sales number counts as the most significant factor as a critical driver for forecasting accuracy. Statistical approaches for current time series forecasting provide simple and straightforward methods. However, utilizing machine learning methods (ML) and especially Neural Networks (NNs) can be exploited to improve time-series predictions and achieve accuracy improvements (Makridakis et al., 2018).

This report is organized in the following manner: Section 3 gives the literature review on textual analysis/ text mining for analyzing the free (unstructured) text information in the finance sector, then reviewing extensive literature on how to utilize Artificial Intelligence (AI) or Machine Learning methods (ML) for promoting sales forecasting accuracy. It is followed by Section 4 that discusses the methodology used to implement the textual analysis and AI-based forecasting model to predict corporate performance. Finally, sections 5 and 6 provide the conclusion and discuss the potential value and future scope that textual analytics and algorithmic financial forecasting provide to the organization and its challenges.

3. LITERATURE REVIEW

This research is conducted by reviewing extensive and comprehensive literature, including analyzing cases where Artificial Intelligence and big data techniques have been implemented in corporate finance. The views of various researchers, academics, and others related to data science emerging with finance activities will be collected and analyzed. Keywords applied in this research will be used to search online databases such as Google Scholar and other digital libraries and websites, including published works and blog articles. The literature review focuses on two topics: ***1) Text Analytics (or text mining) in financial reporting, and 2) Algorithmic financial forecasting in corporate finance***, to understand the better way to apply data science by implementing best practices in the corporate finance sector.

3.1. Text Mining / Textual Analytics in Financial Reporting (Qualitative Analysis)

Existing analytical models rely heavily on structured numerical data. But within many domains, some of the most useful information is buried within unstructured texts. In corporate finance, ample resources of unstructured data such as news articles, finance forum, earnings calls transcripts, conference calls (management presentation and Q&A sections), analyst reports and research notes, and Securities and Exchange Commission (SEC) filings, all provide valuable and meaningful information for evaluating company performance. Fortunately, text analytics can bridge the gap. Text mining or text analytics is the process of using technology to derive useful information from text data. Textual analysis is the process of examining the content, structure, and functions of messages contained in texts based on linguistic theory. The main goal of textual analysis is to extract insightful information from textual resources.

In recent years, several textual analyses for financial reports are trying to quantify the textual data into meaningful predictors for the company's future financial performance (Nguyen & Huynh, 2020). In the early stage of financial textual analysis, people deployed keyword searches and tried to quantify annual reports (10-K) disclosure attributes (Liu et al., 2020). ***Keyword matches, total word count, and sentence count*** could provide initial descriptive analysis on business performance or risk assessment (Liu et al., 2020). Most recently, Bach et al. studied text mining for big-data analysis in finance (Pejić Bach et al., 2019). After Bach et al., Gupta, A., Dengre, V., Kheruwala, H.A. et al. (2020) performed a comprehensive review of text-mining applications specifically in corporate finance, banking, and financial forecasting. Gupta et al. investigated that advanced techniques and algorithms were used on some text-mining applications research for corporate finance as early as 2016 (Gupta et al., 2020, p. 19). For example, Guo, L., Shi, F., & Tu, J. et al. (2016) implemented text-mining algorithms by merging the News Analytics database and the Thomson Reuters News Archive database to conduct sentiment analysis. Three algorithms,

namely ***Naive Bayes (NB)***[3]***, Support Vector Machine (SVM)***[4]***, and Neural Networks (NNs)***[5], were run on the merged dataset. Neural Networks (NNs) produced the highest accuracy rate with a 79.6% outcome within these three algorithms. Guo et al. (2016) concluded that neural networks outperform many other machine learning techniques in classifying news into categories and can be used for text mining-based finance studies (Guo et al., 2016, p. 15).

Lewis and Young (2019) discussed the importance of text mining in financial reports. However, they preferred ***Natural Language Processing (NLP)***[6] methods with the following two advantages than other approaches: 1) NLP has a more powerful capability to deal with immense amounts of data and prevent data overload throughout automated procedures; 2) NLP is able to identify the underlying important latent features (Lewis & Young, 2019; Gupta et al., 2020). Lee et al. (2018) also implemented ***clustering, sentiment analysis*** to evaluate parameters such as correlations between text patterns (positive or negative tone in the business text) in the company's reports and sales performance. Lee's (2018) study focused on identifying text patterns for financial performance by retrieving annual reports (10-K) of US-listed companies (Lee et al., 2018; Gupta et al., 2020). As of the most recent research, there is still little work done to form the predictors using the textual data from financial reports to assess enterprise survival scores or corporate bankruptcy risk. Nguyen and Huynh (2020) proposed a novel financial ***dictionary-***

[3] **Naive Bayes (NB)** is a common machine learning algorithm often used for text classification. The Naive Bayes classifier is a supervised learning model, where a human helps to train pattern-detecting. Naive Bayes classifier uses probabilities of events to make predictions for the purpose of classification.

[4] **Support Vector Machines (SVMs)** are a set of supervised learning methods which learn from the dataset and can be used for both regression and classification. It is a vector space based machine learning method where the goal is to find a decision boundary between two classes that is maximally far from any point in the training data.

[5] **Neural Networks (NNs)** are a series of algorithms that mimic the operations of a human brain to recognize relationships between vast amounts of data. They are used in a variety of applications in financial services.

[6] **Natural Language Processing (NLP)** is a field of artificial intelligence that enables computers to analyze and understand human language. It is an unsupervised learning model, where the computer finds patterns in text with little human intervention.

based sentiment classifier to analyze the management tone reflected through the company filings to SEC. Specifically, they compare the predictive performance of traditional Altman Z-score[7] based models (Altman et al., 2017; Altman and Sabato, 2007) with the ones trained on new text features (especially those based on textual data with their sentiments) from textual data of SEC filings. Nguyen et al. (2020) performed the experiments and concluded that the new model and the combined model built on both types of data (quantitative features based on Altman Z-score + qualitative text features from SEC filings textual data) could significantly improve model fitness and model predictivity power. In their experiments, they used the Altman factors as the baseline model. Then they developed the text and combined model by forming textual features via a counting mechanism (specifically using counting sentiment words and counting sentiment sentences) and financial dictionary-based sentiment classifier. Nguyen et al. (2020) used the proposed model to investigate the relationships of the textual features with the corporate probability of default. They also examined predictivity powers on both k-fold cross-validation and one-year-ahead prediction.

As early as 2008, Li (2008) published his well-known research regarding the annual report readability and earnings persistence. In his study, Li (2008) examined the link between annual report (10-K) readability and firm performance for a meaningful sample in 2008 (Li, 2008; Loughran & McDonald, 2016). He then found that firms with lower reported earnings tend to have annual reports that are harder to read (i.e., high Fog Index[8] values or high word counts). After Li's (2008) research, another pioneering works of Loughran and McDonald ((2011) (2016)) built a dictionary with six wordlists[9] associated with each sentiment category (Negative, Positive,

[7] The **Altman Z-score** is the output of a credit-strength test that gauges a publicly-traded manufacturing company's likelihood of bankruptcy. It uses five financial ratios that can be calculated from data on a company's annual 10-K report to assess if there is a high probability of becoming insolvent.

[8] **Fog Index** is one of textual analytics designed to indicate how difficult a passage in English is to understand.

[9] Loughran and McDonald created six different word lists in 2011 by examining word usage in at least 5% of 10-K (i.e., annual reports) during 1994-2008. The sentiment lists are based on the most likely interpretation of a word in a business context. The wordlists are updated continuously. The latest version 2018 could be found in the resource link: https://sraf.nd.edu/textual-analysis/resources/

Uncertainty, Litigious, Strong Modal, Weak Modal). Loughran and McDonald (2011) showed that some negative words in the Harvard Dictionary are actually neutral or even positive in the financial context, such as depreciation, liability, and so forth. The newly developed wordlists offer a more accurate tone for financial text than the traditional Harvard Dictionary (Loughran & McDonald, 2016; Nguyen & Huynh, 2020). Nguyen et al. (2020) then applied Loughran and McDonald's sentiment wordlists to propose a novel financial dictionary-based sentiment classifier and a new textual features model.

In addition to the above research, there is still a need to develop a system that could perform text-mining techniques on dynamically obtained data to produce real-time output results and enable even better insights. Gupta et al. (2020) concluded that leveraging text-mining techniques and financial data analytics for corporate finance applications can produce a model that could be the most efficient model for this domain. The outputs obtained from mining textual data can be integrated with those from numerical financial data. Thereby, providing a model that deliberates on historical data (quantitative analysis) and forward-looking opinions from diverse sources (qualitative analysis) could enhance more powerful predictive analysis.

3.2. Algorithmic Financial Forecasting in Corporate Finance (Quantitative analysis)

As Steve Lucas (2012), former SAP and Salesforce executive, emphasized in his original SAP HANA Blog posting that "every company needs a real-time data platform" by stating:

> *"The Income Statement and Balance Sheet are at best rear-view measures of the top line and bottom-line. They provide a snapshot in time of all that has **happened,** but very little, if any, indication of what is **happening** in the enterprise. (Alles & Gray, 2016; Beyond the Balance Sheet, 2012)"*

This paper considers how to help an enterprise run its business more proactively in forward-looking rather than only with rear-view perspectives. We want to learn how a company will

gain intelligent, actionable insights to identify the potential risk and opportunity during midstream. And Lucas (2012) stated:

> *"What is needed is the ability to ask complex questions going across the volume and variety of data in an **interactive manner**, and most importantly in **real-time.** (Alles & Gray, 2016; Beyond the Balance Sheet, 2012)"*

In an organization, the management team uses budgeting, planning, and forecasting processes to track financial performance and evaluate whether the company is heading in the right direction. Financial forecasting also allows decision-makers to make necessary adjustments and develop strategic business plans. The more accurate the forecasting results, the easier it is to gain investors' trust to stimulate stock price and market capitalization. The financial forecasting process seeks to produce insightful and accurate future revenues and expenditures patterns. The identified patterns help the company guide policies, pragmatic decision-making, cash-flow management, and strategic planning (Neevista Pty Ltd | Medium, n.d.).

As companies grow in terms of technology and customer-oriented, Kumar, Shankar, and Aljohani (2020) stated that an accurate and efficient forecasting model (all-inclusive demand, supply, and price forecasting models) provides the company a better position to handle customer needs. Thus, an accurate forecasting model has a direct impact on customer satisfaction (demand forecasting) and inventory stock-out (supply forecasting) (Kumar et al., 2020). We are living in an era of big data, and companies are collecting data from multiple dimensions. Kumar et al. (2020) suggested companies moving from traditional forecasting techniques (numerical statistics approach) towards advanced data science methods (merged with unstructured and real-time data approach). Regardless of how, it is further complicated by the effects of seasonality, promotions, and product proliferation (Lu & Wang, 2010).

3.2.1 Big-Data Demand-Driven Forecasting

Kumar et al. (2020) aimed to improve demand forecast accuracy via developing a big-

data demand-driven forecasting model. Typically, fiscal budgeting and forecasting start from sales prediction. Estimates of demand are the foundation of all planning activities because they directly impact the sale forecast, referred to as top-line performance. To achieve this goal, Kumar et al. (2020) studied a ***back-propagation neural network***[10]-based model trained by ***fuzzy inputs*** and compared it with benchmark forecasting methods on time series data. They used historical demand and sales data in combination with advertising effectiveness, marketing expenditure, promotional campaigns, and marketing events. To be more specific, this proposed demand-driven forecasting model follows the five processes presented by Chase Jr (2013) summarized as below:

Demand Translation: Actual and forecasted sales (demand) data are passed to supply-side production line and distribution planning database (supply planning) (Chase, 2013). This step will mitigate the gap between actual demand and actual supply.

Demand Sensing: Sensing customer purchase behaviors and utilizing upstream data within the value chain to generate a more accurate demand forecast, taking into consideration of historical buying patterns and product seasonality (Folinas et al., 2012).

Demand Shaping: Developing an optimized and steady plan of demand and supply so that the sales and profitability targets are met and customer satisfaction metrics are achieved.

Demand shifting**:** Aligning demand patterns and supply capacity through the two types of strategies: 1) at the point of sales: incentivizing customers to buy an alternative if the products initially demanded is not available; 2) at the point of supply: suppliers negotiating with the sales team to shift demand due to capacity constraints (Chase Jr, 2013).

Demand Orchestration: Developing demand plans that ensure the expected trade-offs between demand opportunity and demand risk are optimized (Kumar et al., 2020).

[10] **Back Propagation Neural Network:** Back-propagation is the essence of neural net training. It is the method of fine-tuning the weights of a neural net based on the error rate obtained in the previous iteration. https://www.guru99.com/backpropogation-neural-network.html

Even though past research has shown limited literature on machine learning applied for demand-driven forecasting using big data analytics, these five steps seem reasonable. Kumar et al. (2020) would be the first to develop a big data-driven framework for demand shaping, customer sensing, and enhancing future forecasting accuracy. They used time series model as a baseline and applied ***fuzzy artificial neural networks (fuzzy ANN)***[11] based classifier in a big data environment. In their experiments, Kumar et al. (2020) performed simulation analysis based on various marketing effectiveness plans with the factors such as marketing channels, advertising adstock decay effect, and advertising impacts on sales. They also synchronized the demand and supply data, which gave their proposed model better forecasting results. The main difference for Kumar et al.'s model [Figure 2][12] [Figure 3] compared with past research could be attributed to the following three modifications in their novel forecasting model:

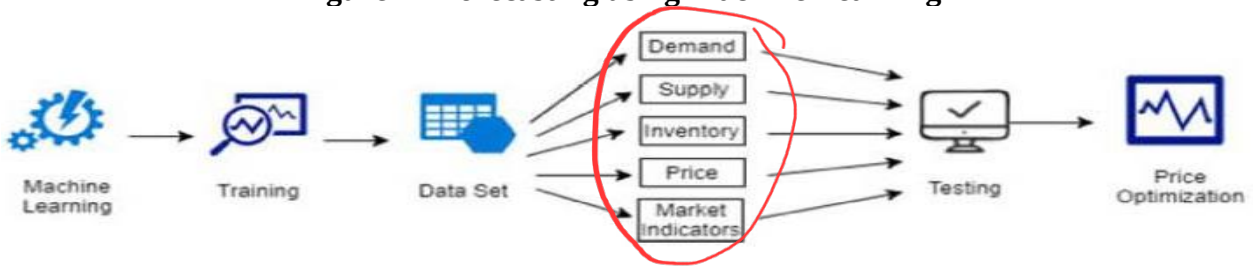
1) The adoption of an efficient and error-free fuzzy classifier resolves the computation power limitation working on big datasets, which requires excellent self-learning capability. The fuzzy neural network was first proposed by Yu & Huarng (2010) and applied in several domains with excellent results (Lolli et al., 2017).
2) Kumar et al. (2020) included promotional marketing activities as predictors combining historical demand and other essential factors. This method enhances the forecasting more aligning with customer purchase behavior and results in more accurate demand forecasting.
3) Kumar et al. (2020) 's model represents an extension of previous demand-driven forecasting work developed by Chase Jr (2013) in terms of modified architecture.

With these above modifications made and examined in Kumar et al.'s (2020) model, their big-data-driven demand forecasting model was concluded to improve financial ***prediction accuracy*** and ***model interpretability***. By adopting an efficient and error-free fuzzy classifier, the results

[11] https://www.tutorialspoint.com/fuzzy_logic/fuzziness_in_neural_networks.htm
[12] https://scholar.harvard.edu/linh/financial-forecasting-using-machine-learning

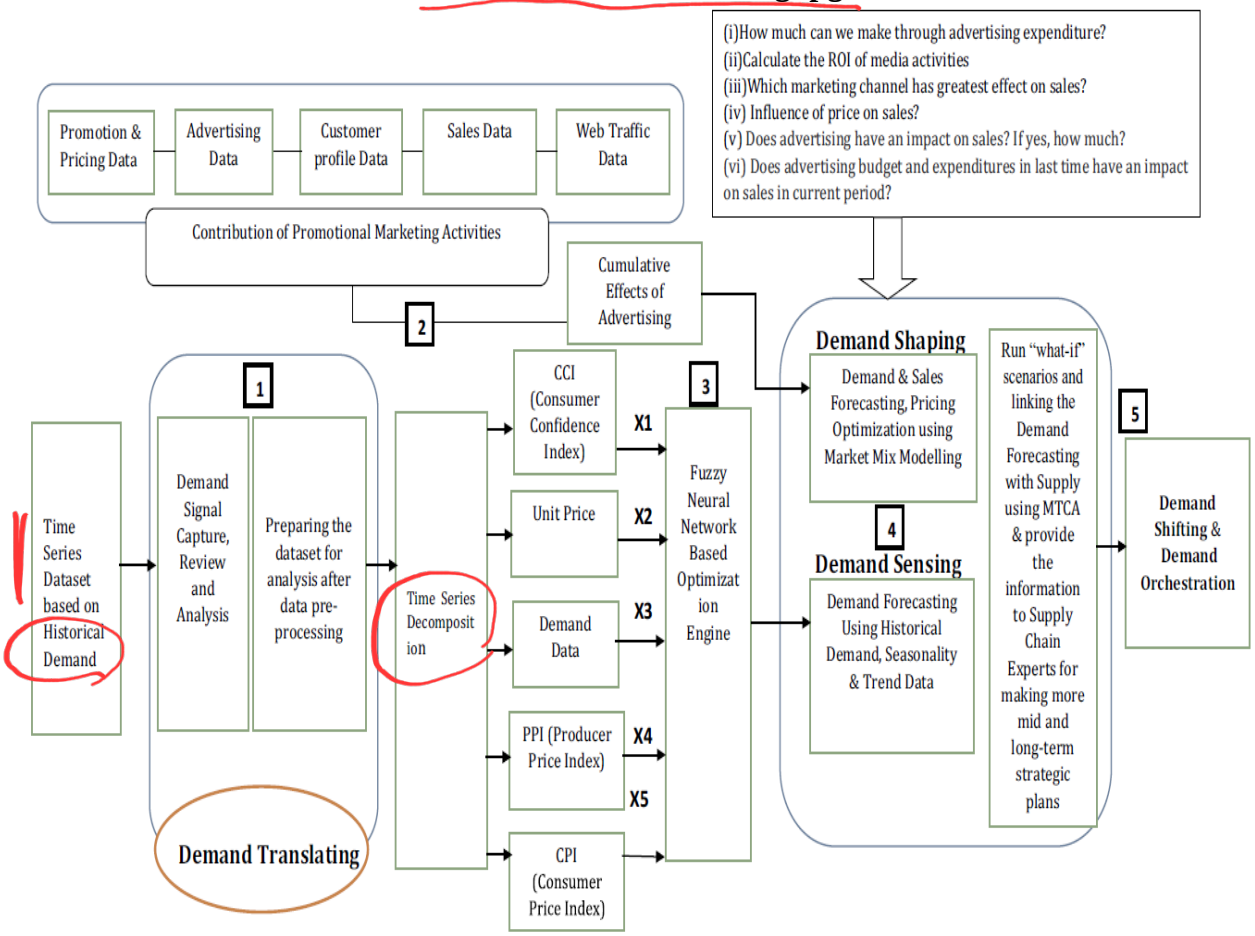
showed a comparatively lower discrepancy between actual and forecast results than all other models[13] [Figure 4].

Figure 2 - Forecasting using Machine Learning



Source: https://scholar.harvard.edu/linh/financial-forecasting-using-machine-learning

Figure 3 - Kumar et al. (2020) 's big data-driven framework for demand-driven forecasting [pg4].



[13] The experiments was applied with TV manufacturing industry dataset, considering five years data as the training data and 2090 observations (80%) are selected for this purpose and the next 523 observations (20%) are used for testing purposes to gain more accurate forecasting results (Kumar et al., 2020, p. 10).

Figure 4 - Performance matrices[14] result on demand data based on Kumar et al. (2020) 's model [pg13]

Model	MSE	MAPE	MAD	Bullwhip effect	Net stock amplification
ARIMA	743.66	40.46	202.21	1.57	1.78
Artificial neural network (ANN)	7.89	9.25	29.36	1.11	0.90
Support vector machine (SVM)	33.26	9.62	48.59	1.31	1.12
Random forest	18.29	11.57	52.37	1.06	0.45
Multiple linear regression (MLR)	117.34	27.36	189.55	1.14	0.63
Fuzzy neural network	6.71	4.49	21.95	0.99	0.37

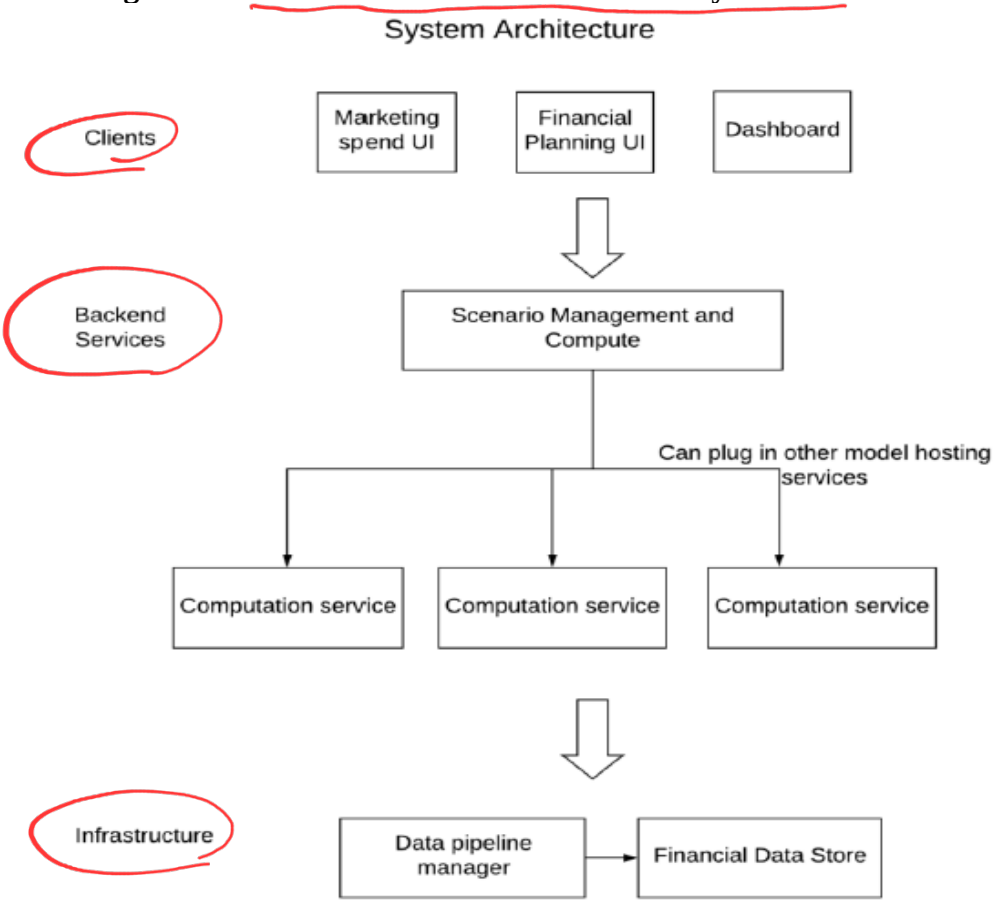
The performance matrices results based on different models compared with Kumar et al.'s model are presented in Figure 4. The smaller the error measures, the better performance of the model. Kumar et al.'s (2020) model presents the lowest error measures and could be a more superior model than the others.

3.2.2 Uber's Financial Intelligence- Financial Cloud Platforms

Most of the research approaches on algorithmic financial forecasting models still stay in theory, while scarce experimental research or real business cases. Uber's ride-sharing business model has grown rapidly with global expansion and business penetration in the past five years. To fulfill the need for quick response in business strategy, Uber Engineering developed an internal shared forecasting infrastructure—Financial Cloud Platforms. This platform improves collaboration and coordination while providing efficiencies in data storage, tool configuration, and knowledge sharing (Alles & Gray, 2016; Lee et al., 2018, p. 4). Uber's financial forecasting system implemented artificial intelligence (AI) algorithms that worked on the model shown below [Figure 5] (Song, 2018) for financial planning. The internally built financial forecasting platform at Uber incorporates multiple layers, from user interface (UI) to computation toward data. The engineers at Uber developed a cloud platform that allows their corporate finance team to scale on time, location, incorporate new updated business models at the same time.

[14] Kumar et al (2020)'s accuray performance matrices based on Error Measures defined as below:
*Mean Square Error (MSE): measures the average squared difference between forecasted and actual values.
*Mean Absolute Percentage Error (MAPE): measures the (absolute) size of each error in percentage terms, then averages all percentages.
*Mean Absolute Deviation (MAD): the average absolute distance between each data point and the mean.
*Bullwhip effect: is caused by fluctuations in data supplied to companies that are further up the supply chain.

Figure 5 - Uber's Financial Cloud Platform System Architecture



Source: https://eng.uber.com/transforming-financial-forecasting-machine-learning/

- **UI Layer – The Top Layer:** This layer consists of three components, including Marketing Spending, Financial Planning, and a Dashboard. This digital back-end system with User interface (UI) enables the financial planning team to do the budget allocation (Corporate finance) and scenario planning (city teams), collaborating concurrently in real-time, all running on the same platform (Shivang Blog, 2018).
- **Backend Services – The Middle Layer:** This layer holds the business logic, and it is the Scenario Optimization & Computation platform. It is like a big While Loop iterating over tons of data sets. Through millions of iterations with different financial metrics, it creates scenarios for each city like the spending scenario. It then computes the metric in these scenarios to get the optimal solution (Shivang Blog, 2018).

- **Infrastructure Layer – The Last Layer:** This layer includes the data pipeline, financial data warehouse, and metric store. It is the layer serving the data to the platform.

Traditionally, companies plan annual or quarterly forecast planning to fine-tune or adjust business strategy. Uber plans differently with realizing the fast-growing model of its business model. Annual or quarterly planning is ineffective and inefficient. The Uber ***finance cloud platform*** enables their corporate finance team to continually adjust financial planning throughout the year rather than relying on a quarterly or annual schedule (Song, 2018). This AI-driven platform helps the management team react to trend exploration and initiate immediate action plans.

3.2.3 Foreseen Modernized Forecasting

Though there are a couple of top companies in B2B providing AI-driven software or platform to transform the way in conducting sales forecasting, there is a need to comprehensively integrate the sales forecasting model with other corporate finance forecasting platforms. A classic example would be the Salesforce Einstein Forecasting model, which is among the most reliable and powerful self-learning sales forecasting channels that leverage the company's data. This AI-powered sales forecasting tool can instantly deliver highly accurate forecasts on Salesforce CRM for sales (Ghosh, 2019). However, without comprehensively integrated with supply chain planning, marketing efficiency, and resource efficacy, companies could not enhance profitability thought only focus on revenues.

Annor, Albert & A., Ayman & Chunting, Yang. (2019) illustrated that AI drives digital platforms and makes it possible to easily integrate with other technologies and reinforce each other to develop efficient and outstanding forecasts (Antwi et al., 2019, p. 6). Big 4 CPA firms recently put more emphasis on developing algorithmic forecasting consultant services. With algorithmic forecasting, Finance talent does more insight-driven work and less manual dull work. Instead of spending time grinding through spreadsheets, humans get to bring their expert

judgment to the process (Hogan & Merrill, n.d.). AI-based forecast methodology must be intelligent, agile, and able to reflect market dynamics accurately. The corporate finance talent model should also evolve to keep up with tech change; this may require finance professionals to acquire a diverse mix of skill sets than they have in the current day. Algorithmic forecasting is a transparent way to help improve the forecasting process, and it depends on collaboration among finance, data analytics, and business teams. To bridge the gap between business strategy and modern technology, tech-savvy finance professionals could drive the integration.

4. METHODOLOGY

4.1. Textual Analytics (Qualitative analysis)

Although textual analysis is a useful tool, its value multiples when paired with data science techniques. The rapid growth in computer processing capability increased big data availability, and the development of new software tools has stimulated the usage of textual analysis in many disciplines. This study utilized text analysis software **RapidMiner** to apply in two retail industry companies that were performing at different levels. With this methodology, we could compare the different levels of ***passive verbs, poor readability, and negative sentiment*** of each company and anticipate future financial performance. This method could be a benchmark analysis through reading a competitor's published financial statement disclosure, annual or quarterly reports, MD&A, and social media posts regarding the firms' financial performance.

4.1.1 Three Text-Mining Methodologies Used in Corporate Finance

(1) Word clouds and keyword extraction: Word counts (frequency) can be used to list the most frequently occurring concepts or words in a given text. That can be useful in the corporate finance sector for several examples as following:

E.g., If the word 'delivery' appears most often in a customer contract, this might suggest there are issues with a company's delivery service and revenue recognition.

E.g., If the word 'consignment' frequently appears in a customer contract, this might suggest further investigating the appropriateness of revenue recognition.

E.g., Suppose the word 'ownership' frequently appears in a lease contract. In that case, this might suggest that capital lease accounting should be considered. Further investigation would be necessary if the company intends to undermine profit through operating lease manipulation to reduce the tax obligation.

From these three examples shown above, a financial analyst could learn any concerns on revenue recognition. The top management team would intend to set sales performance over-positive to boost the stock price or undermine profit to avoid tax burden. Text analytics could help detect if any concerns of revenue manipulation to secure budgeting and forecasting baseline.

(2) Readability Measure:

(2.1) Fog Index (Li, 2008; Loughran and McDonald, 2014): This index is designed to indicate how difficult a passage is to understand. A Linear combination of average sentence length and proportion of complex words is comprehensively considered to measure the fog index. Namely, Fog Index = 0.4 (average number of words per sentence + percent of complex words) (Li, 2008; (Loughran & McDonald, 2016, p. 8). Based on this readability measure, Fog Index reflects the number of years of education needed to understand the text on a first reading [Figure 6] (Loughran & McDonald, 2016, p. 8). Generally, Fog Index between 10 and 14 will be ideal and acceptable. Financial report readability can be measured by Fog Index based on the idea that managers can obscure their financial reports by making them harder to read. A larger Fog Index indicates lower financial report quality. Same as lease contracts and customer contracts, the many complex words involved in contracts would make users hard to understand. Finding the most frequent appearing

complexity words is critical to show the readers which words would have a disproportionate impact on their understanding.

However, Loughran and McDonald (2014) also pointed out there are limitations of the Fog Index. First, the index assumes a word to be complex based on the number of its syllables. For example, the term "interesting" comprises four syllables and is therefore considered complex by the Fog Index. Generally, it is hardly considered a problematic word. Secondly, how to define "readability" for business documents? The specific nature of terms used in financial statements or specific sentiment words related to financial statements is different from a generic fog index. Loughran and McDonald (2014) empirically demonstrated that the Fog Index would be a poorly specified readability measure when applied to business documents (Guo et al., 2016). For example, words such as "management," "profitability," and "operation" are presumably easy for business decision-makers to comprehend while would be classified into complicated words based on Fog Index measurement. Business text commonly contains many multi-syllable words used to describe business operations, which typically result in a more difficult readability score using the Fog Index. In the implementation analysis ***4.1.2 section***, an ***uncertainty*** wordlist developed by Loughran and McDonald (2018 updated) (Dame, n.d.-b) is applied to compare the number and type of uncertainty words associated with each 10-Q.

Figure 6- Fog Index

Fog Index	Reading Level by Grade
20+	Post-graduate plus
17–20	Post-graduate
16	College senior
15, 14, 13	College junior, sophomore, freshman
11–12	High school senior, junior
10	High school sophomore
9	High school freshman
8	8th grade
7	7th grade
6	6th grade

Source: https://www.researchgate.net/figure/Gunnings-fog-index-level-31_tbl1_344162323

(2.2) Firm-Level Complexity (Loughran and McDonald 2020) (Dame, n.d.-a) – Most recently, Loughran and McDonald (2020) created a list of 374 words that proxy for the complexity of various firm attributes. This measure's essence is that higher distinct occurrences of words associated with complex business events, transactions, and intricate business practices should be linked with greater levels of firm-level complexity. Thus, firm-level discussion or executive summary in the annual report (10-K) / quarterly report (10-Q) relating to accounting terms (accrue, carryforwards, and leaseback), corporate operational strategies (bankruptcies, partnership, and restructure), M&A activities (acquired, merger, and takeovers), legal issues (lawsuit, litigation, and contract), international operation management (foreign, global, and worldwide), financial leverage (derivatives, hedge, and unexercised), and intangibles (patents, trademarks, and copyrights) are included in Loughran and McDonald (2020) 's new created word list [TABLE 1]. In the implementation analysis ***4.1.2 section***, a ***ComplexityWords*** wordlist developed by Loughran and McDonald (2020) (Dame, n.d.-a) is applied to compare the number and type of complexity words associated with each 10-Q.

(3) Sentiment Analysis / Tone: Sentiment analysis for customer feedback and customer support provides customer satisfaction information. Learning how customers perceive a product or service could help project sales returns provision and reevaluate revenue recognition—the more negative the tone, the more concerns of the sales prediction. Sentiment analysis for lease contracts can be examined by checking the use of uncertainty words (e.g., approximate, contingency, uncertain, and indefinite) and weak modal wordlist developed by Loughran and McDonald (2018 updated). The contract ambiguity would make it hard to understand and avoid negotiations deadlock, which would cause misunderstandings, consequences, and unnecessary grievances. Sentiment analysis can apply to contracts analysis for customer contracts and lease agreements to mitigate legal risk. Through utilizing advanced linguistic and AI capabilities, users could speed up discovering, reviewing, and analyzing workflows by pinpointing contract clauses. In the

implementation analysis ***4.1.2 section***, I applied the **VADER** sentiment analysis extension of **RapidMinder** to determine the tone (positive versus negative) of the 10-Qs/10-Ks and generate polarity scores. Polarity scores measure the tone's strength from a +1 (strongest positive tone) to -1 (strongest negative tone). **VADER (Valence Aware Dictionary for Sentiment Reasoning)** is a model used for text sentiment analysis that is sensitive to both polarity (positive/negative) and intensity (strength) of emotion. It is ***Natural Language*** based package and produces sentiment scores based on a dictionary of words (Hutto, 2014/2021). In this paper, the 10-Qs/ 10-Ks Corpus is used to conduct the experiments. Section 7, "Management's Discussion and Analysis of financial conditions and results of operations" (MD&A), is the only used source for sentiment analysis since this section contains the essential forward-looking statements about the companies (Wang et al., 2013, p. 3).

4.1.2. Real Case Implementation via RapidMiner Analysis for Two Companies

This study used data mining methodology and **RapidMiner** Textual Analysis software, a ready-to-use Machine Learning Data Science Platform, to investigate the relationship between the sales performance and the text patterns retrieved from annual reports. RapidMiner is an excellent tool for non-programmers and especially suitable for text mining analysis. To better find hidden patterns in the dataset, RapidMiner coped with Decision Trees statistical modeling technique to build the analysis process. Although annual reports essentially cover past and present financial information with limited future expectation insights, analyzing text content from the latest annual reports and quarterly reports can still help identify patterns or sentences that indicate the company's expected future business performance. This methodology could be applied by corporate finance to understand significant competitors' business trending and help them initiate business strategy. To keep a competitive advantage, a company also must learn how the major competitors react to the microeconomic environment and how they position their near-future business more positive (aggressive) or prone to negative (conservative).

(1) Brief justification for why companies were selected to analyze:

In this study, the following assumptions must be taken into consideration: 1) macroeconomic environment is typical without extraordinary events; 2) annual/ quarterly reports published by US-listed companies which were audited/ assured by Certified Public Accountant were fairly disclosure, without manipulation intent. After considering these two assumptions, the latest company performance in the year 2020 was not selected to perform the implementation to avoid the extraordinary impact of COVID-19. Alternatively, two US public companies in the same industry performing at different levels - Sears Holdings (SHLD) and Target Company (TGT) - were chosen with the year 2017 and 2018 quarterly reports to proceed with the text analytics implementation. These two companies were both broad retailers but had drastically different performance outcomes following quarter three of 2017, which was analyzed. Sears Holdings is the parent company to both Sears and K-Mart, both direct competitors to Target. 2017 was chosen to examine because soon after, Sears Holdings (SHLD) filed for Chapter 11 bankruptcy on October 15, 2018, whereas Target (TGT) was outperforming in a poor retail environment and proving that it is America's new department store. Therefore, it was expected that much of the textual analysis would be different between the two companies. Under the assumption that companies with poor readability, passive verbs, and negative sentiment are more likely to underperform in future periods, it was expected that Sears Holding would do worse in the textual analyses because it was already known that the future performance was very poor.

(2) Data Collection

These two companies are listed public enterprises and have the responsibility to file their periodic financial reports via the Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system with the United States Securities and Exchanges Commission (SEC). Annual reports (10-Ks) and quarterly reports (10-Qs) data set, along with their Management Discussion & Analysis (MD&A), are available in the SEC public EDGAR system to retrieve for processing textual

analysis for this study.

In the first step, I downloaded the raw filings 10-Qs and 10-Ks filings from the SEC-EDGAR website to start the analytical procedure. Through the data gathering process, I collected the fiscal year 2017 10-K filings (2017-Q4) and other three quarters 10-Q filings (2017-Q3; 2018-Q1; 2018-Q2) for each company respectively in HTML format, then converted these data into plain textual content by removing HTML tags, Tables, Figures, attachments, graphs, and other redundant elements from the original HTML files. Secondly, I cleaned and parsed the textual data by removing taggers and stop words, thus putting plain text into a **word vector** [Figure 7]. Finally, I imported the converted files into the RapidMiner Repository database for further analysis [Figure 8]. An essential parameter is the vector creation used in *Process Documents from Files Operators*. The selected vector creation method is ***TFIDF (Term Frequency-Inverse Document Frequency)*** is a term weighting method. It gives higher weight to terms that frequently appear in the document but not many times in other papers [Figure 9]. Operator *"Process Documents from Files"* performs text processing and prepares the text data to apply data mining techniques. Through "*Process Documents from Files,"* the operator reads data from a collection of text files and manipulates it using text processing algorithms. The subprocess transforms the text data into a format that can be analyzed efficiently using conventional data mining techniques such as association mining and cluster modeling (Ertek & Ertek, 2017, p. 27).

Figure 7- RapidMiner: Process Documents from Files Operators

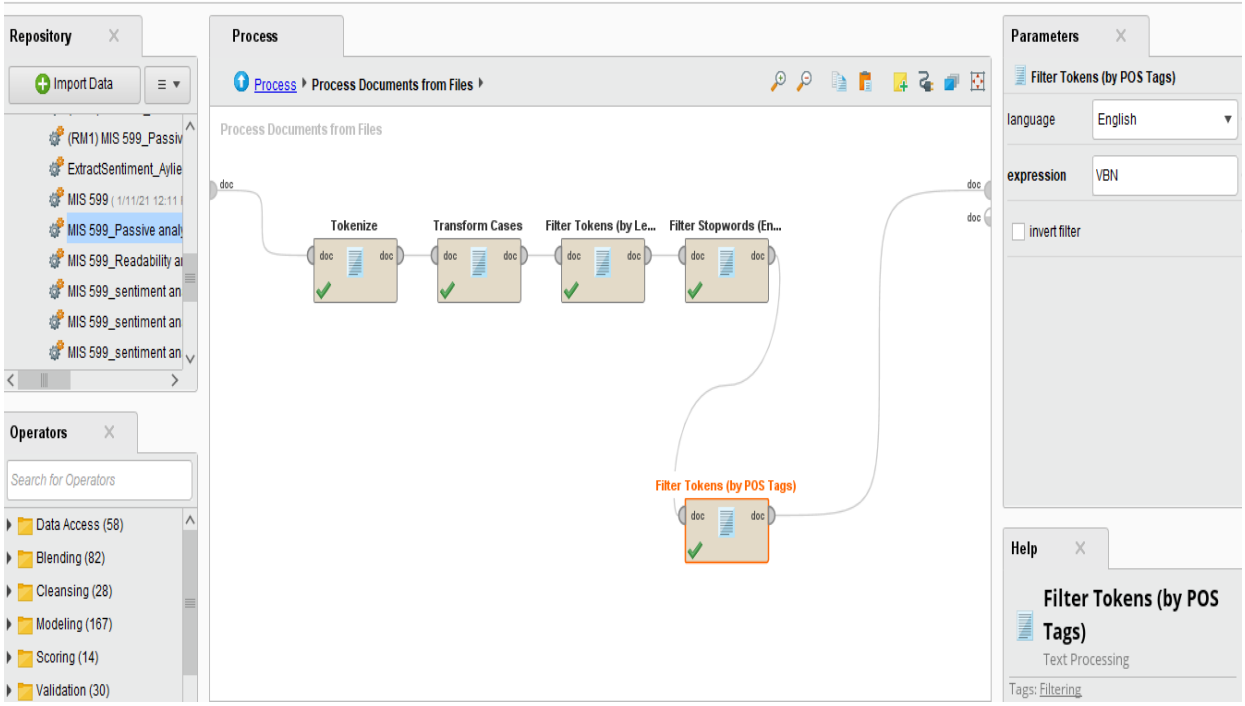


Figure 8- RapidMiner: Edit Parameters: text directories

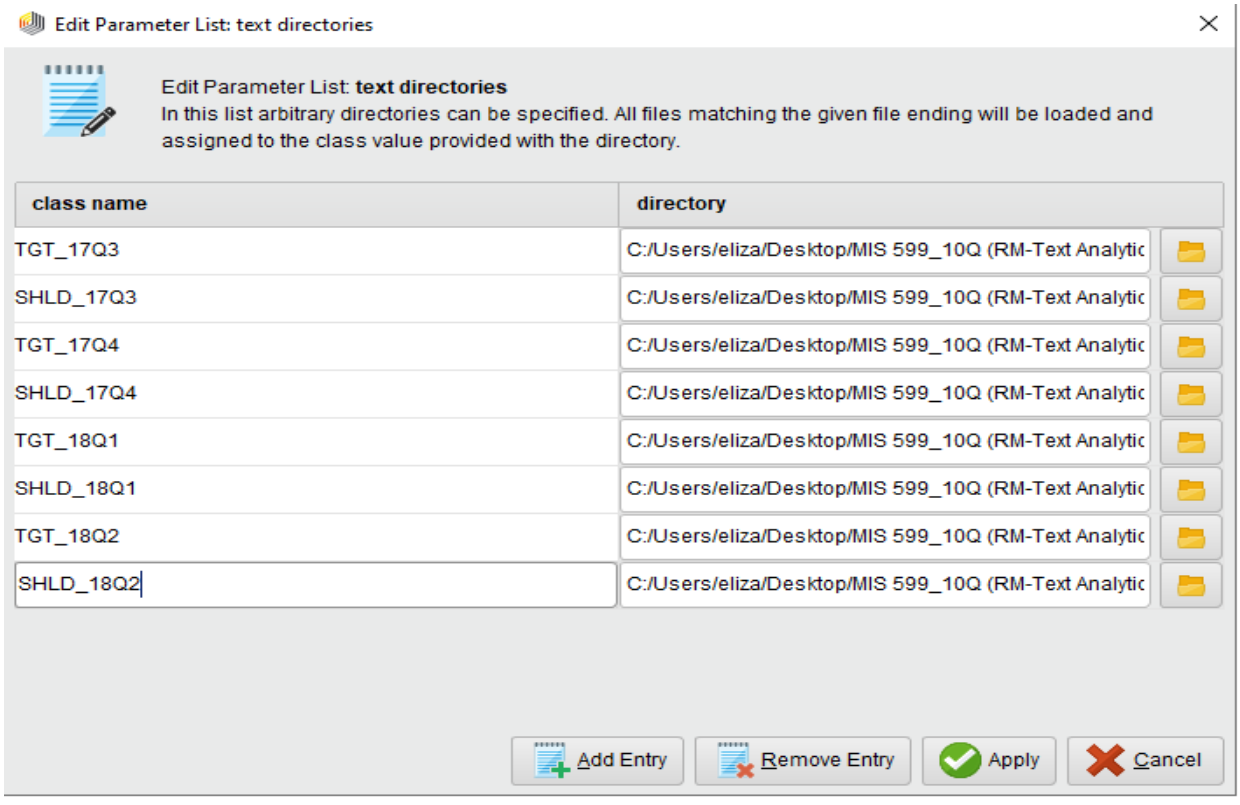
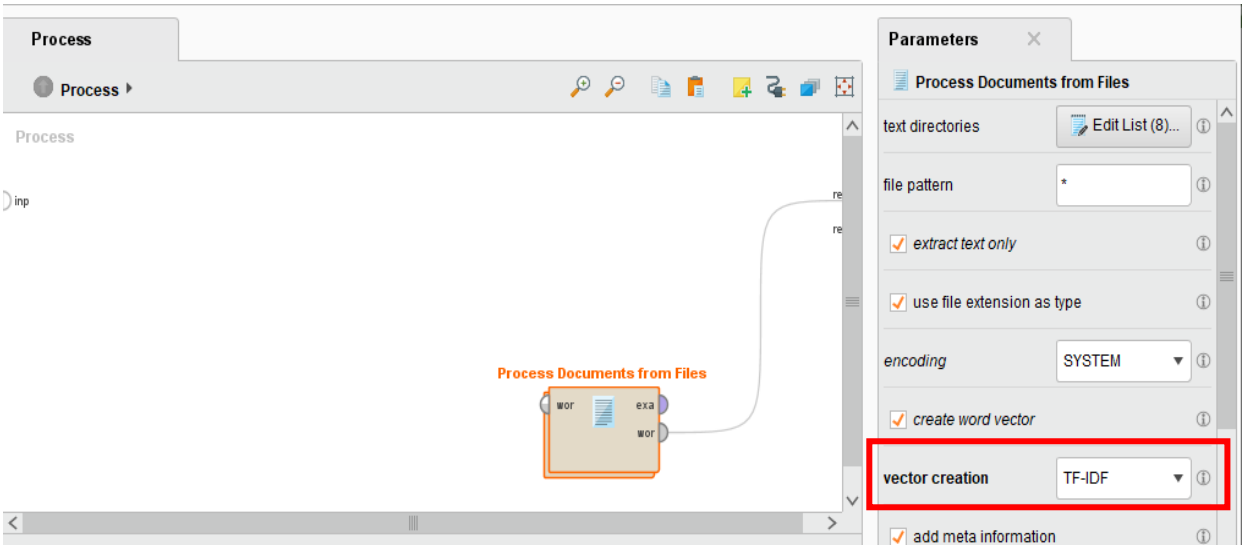


Figure 9- RapidMiner: Edit Parameters: vector creation

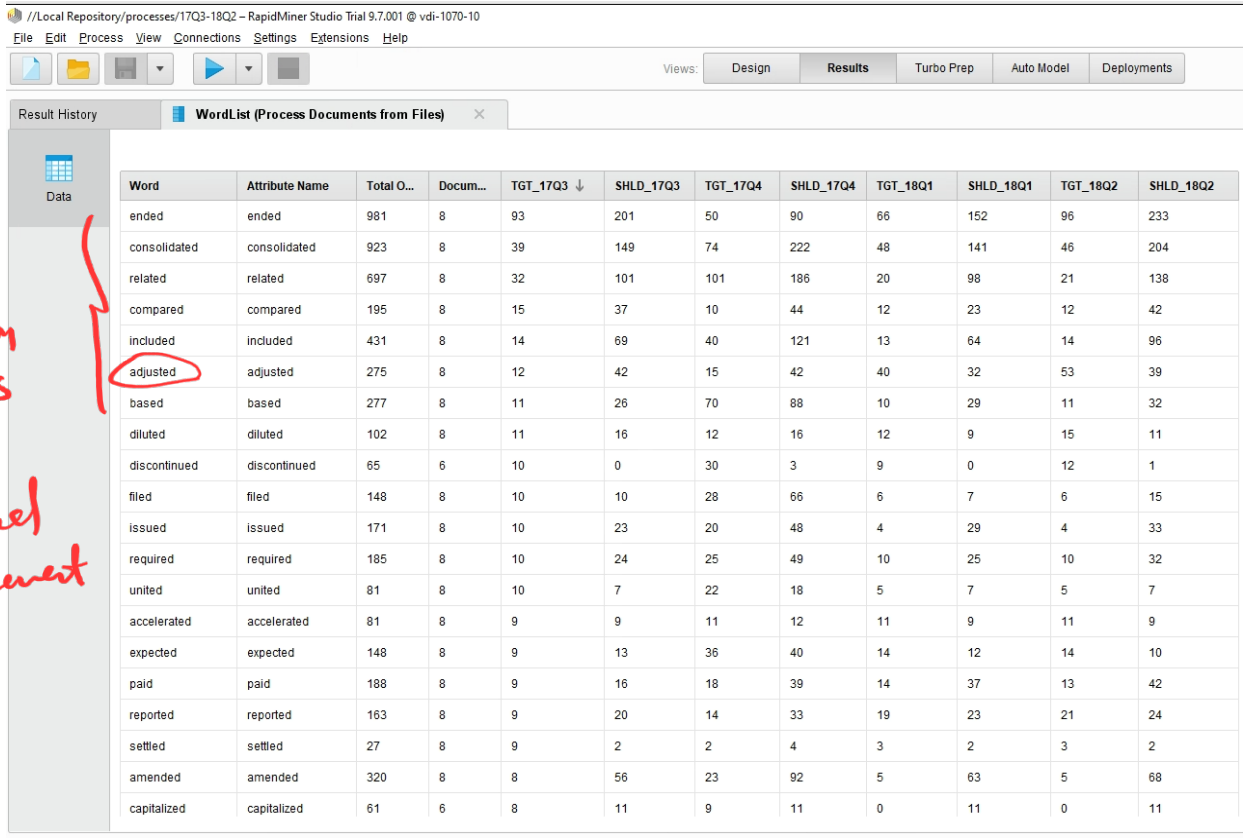


(3) Empirical analysis

(3.1): Passive Verbs Analysis

In contrast with active voice, passive voice is not that persuasive and is the reverse concept. The passive voice could show uncertainty or lacking confidence in the content. Passive verb analysis could indicate an attempt to obscure one's responsibility from the sentence content. In other words, overuse of passive verbs may suggest an attempt to avoid responsibility for what is being stated in annual or quarterly reports. I used the operator *"Filter Tokens (by POS Tag)"* to look for passive tone. The eight lists in Figure 10 below show the top passive words for both Sears and Target from the year 2017 quarter three to 2018 quarter two [Figure 10]. These lists show that Sears used many more passive words than Target. Sears's top passive verb in 2017 quarter three was at 201 with a word "ended," which was almost twice as much as Target's top verb at 93. Sears's top passive verb in 2018 quarter two was at 233, which was more than twice as much as Target's top verb at 96. This finding could signal that Sears has a higher probability of conducting fraud and attempting to obscure their responsibility from financial statements or reporting. Also, this could indicate poor past or future performance since management is separating itself from the actions of the company.

Figure 10- RapidMiner: Passive Verbs Analysis



Word	Attribute Name	Total O...	Docum...	TGT_17Q3	SHLD_17Q3	TGT_17Q4	SHLD_17Q4	TGT_18Q1	SHLD_18Q1	TGT_18Q2	SHLD_18Q2
ended	ended	981	8	93	201	50	90	66	152	96	233
consolidated	consolidated	923	8	39	149	74	222	48	141	46	204
related	related	697	8	32	101	101	186	20	98	21	138
compared	compared	195	8	15	37	10	44	12	23	12	42
included	included	431	8	14	69	40	121	13	64	14	96
adjusted	adjusted	275	8	12	42	15	42	40	32	53	39
based	based	277	8	11	26	70	88	10	29	11	32
diluted	diluted	102	8	11	16	12	16	12	9	15	11
discontinued	discontinued	65	6	10	0	30	3	9	0	12	1
filed	filed	148	8	10	10	28	66	6	7	6	15
issued	issued	171	8	10	23	20	48	4	29	4	33
required	required	185	8	10	24	25	49	10	25	10	32
united	united	81	8	10	7	22	18	5	7	5	7
accelerated	accelerated	81	8	9	9	11	12	11	9	11	9
expected	expected	148	8	9	13	36	40	14	12	14	10
paid	paid	188	8	9	16	18	39	14	37	13	42
reported	reported	163	8	9	20	14	33	19	23	21	24
settled	settled	27	8	9	2	2	4	3	2	3	2
amended	amended	320	8	8	56	23	92	5	63	5	68
capitalized	capitalized	61	6	8	11	9	11	0	11	0	11

(3.2): Readability Analysis

(3.2.1): Uncertainty Words Analysis to Identify Readability using Loughran-McDonald Sentiment Word List.

Secondly, I performed an analysis in RapidMiner to identify data readability using the Loughran-McDonald Sentiment Word List. In *"Process Document from Files Operator*," *"Filter Tokens (by Content) Operator"* was added, and the uncertainty word list was added in "regular expressions" under parameters [Figure 11&12]. The results show that uncertainty words were used much more often in Sears' financial releases than in Target's. For instance, Sears used the term "approximately" 74 times in the year 2017 quarter three. It was the highest-frequently-used uncertainty word on its list in this period [Figure 13]. While Target's highest-frequently-used uncertainty word in the year 2017 quarter three was "risk," which was used only nine times [Figure 14]. Uncertainty words can be used to convolute writing and make bad results or announcements seem less impactful. Since Sears has a much higher count of uncertainty words,

the company is more likely to be attempting to hide financial struggles than Target and may thus be more susceptible to fraud. I applied continuous four quarters data from 2017 quarter three to 2018 quarter two to compare the financial results to see one-year-ahead prediction performance via textual financial information. I then traced back the financial performance one-year-ahead before Sears' bankruptcy in the year 2018 quarter three [Figure 15]. In figure 15, the results align with the original anticipation that uncertainty words were used much more often in Sears' financial releases than in Target's one-year-ahead before Sears bankruptcy.

Figure 11- RapidMiner: Filter Tokens (by Content) Operator

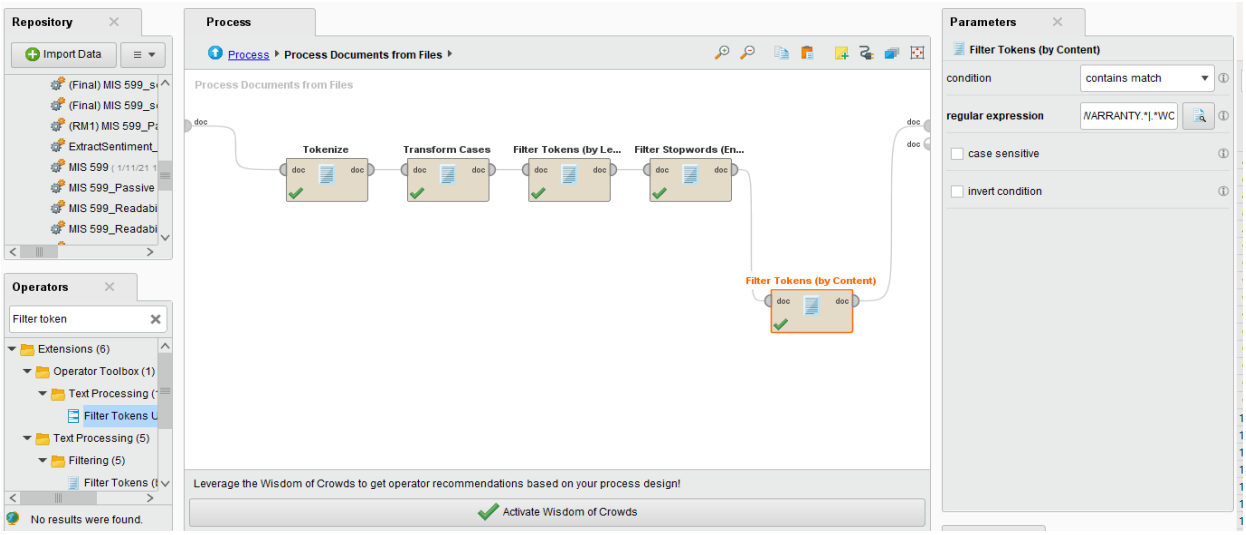


Figure 12- RapidMiner: Edit Regular Expressions

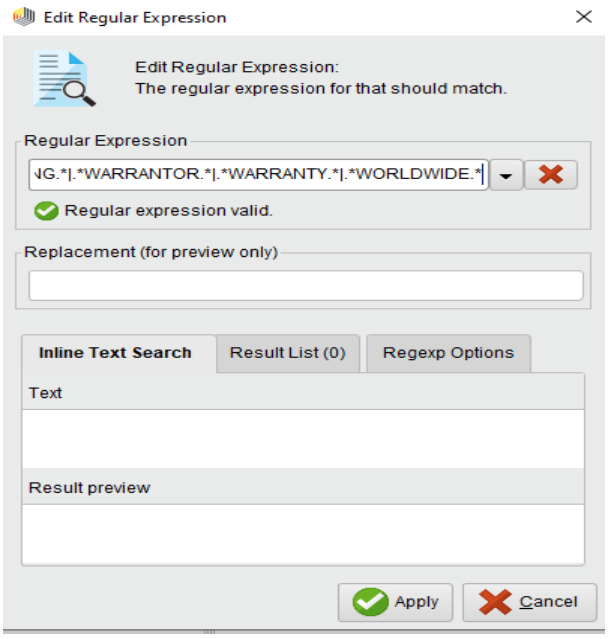
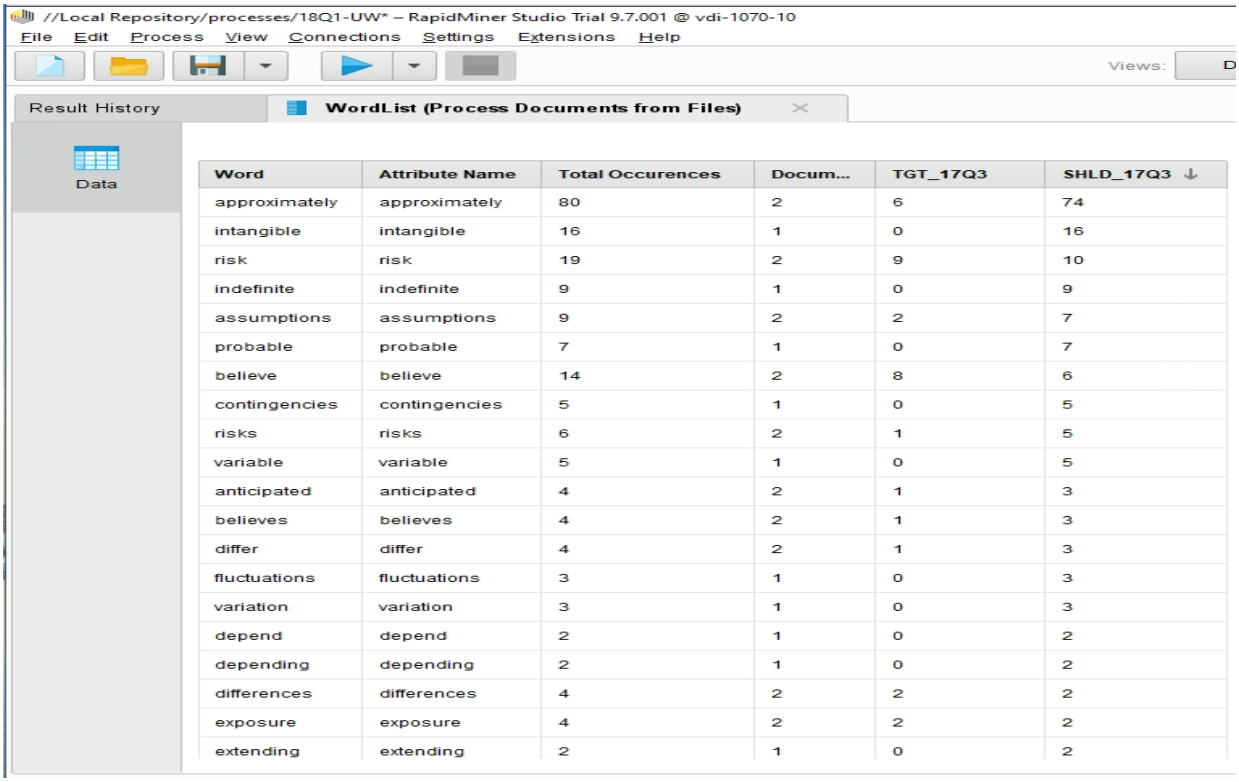
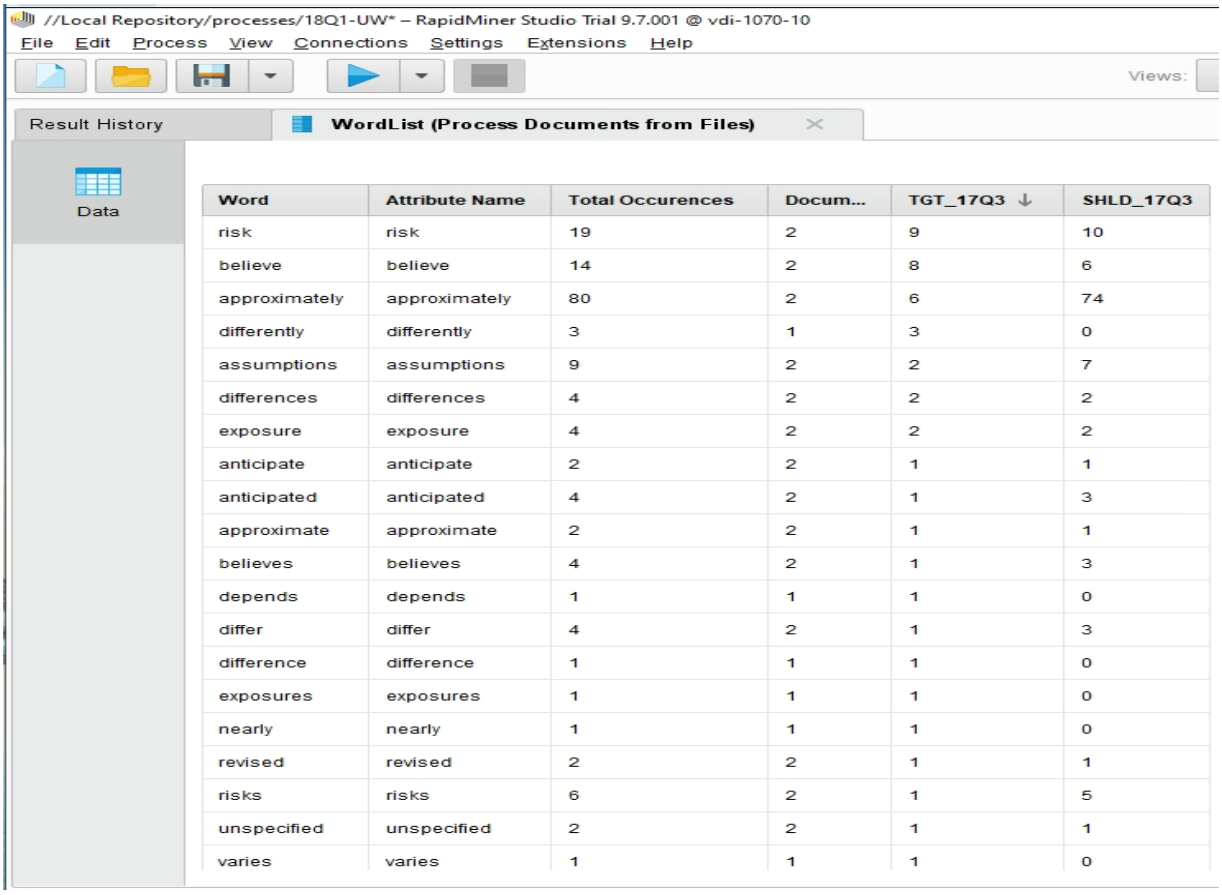


Figure 13- RapidMiner: Uncertainty Words analysis (Sears)



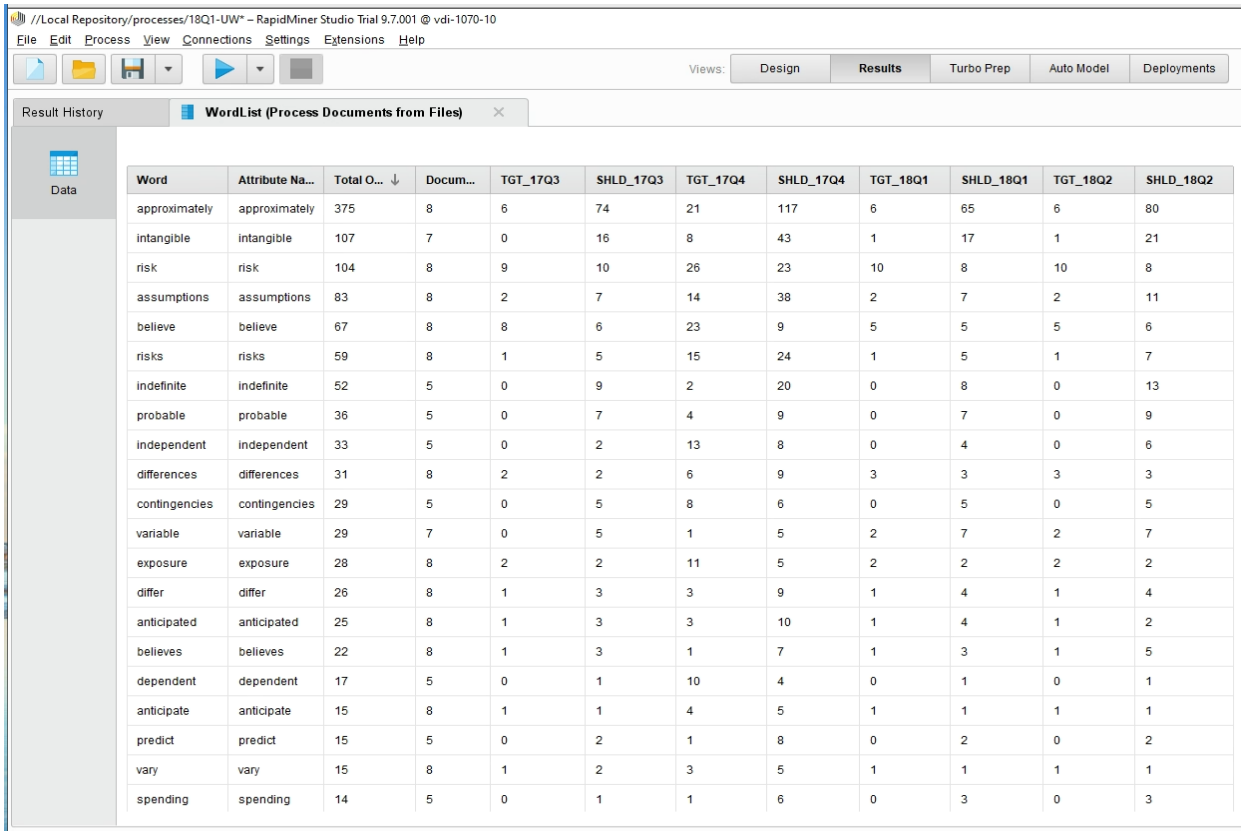
Word	Attribute Name	Total Occurences	Docum...	TGT_17Q3	SHLD_17Q3
approximately	approximately	80	2	6	74
intangible	intangible	16	1	0	16
risk	risk	19	2	9	10
indefinite	indefinite	9	1	0	9
assumptions	assumptions	9	2	2	7
probable	probable	7	1	0	7
believe	believe	14	2	8	6
contingencies	contingencies	5	1	0	5
risks	risks	6	2	1	5
variable	variable	5	1	0	5
anticipated	anticipated	4	2	1	3
believes	believes	4	2	1	3
differ	differ	4	2	1	3
fluctuations	fluctuations	3	1	0	3
variation	variation	3	1	0	3
depend	depend	2	1	0	2
depending	depending	2	1	0	2
differences	differences	4	2	2	2
exposure	exposure	4	2	2	2
extending	extending	2	1	0	2

Figure 14- RapidMiner: Uncertainty Words analysis (Target)



Word	Attribute Name	Total Occurences	Docum...	TGT_17Q3	SHLD_17Q3
risk	risk	19	2	9	10
believe	believe	14	2	8	6
approximately	approximately	80	2	6	74
differently	differently	3	1	3	0
assumptions	assumptions	9	2	2	7
differences	differences	4	2	2	2
exposure	exposure	4	2	2	2
anticipate	anticipate	2	2	1	1
anticipated	anticipated	4	2	1	3
approximate	approximate	2	2	1	1
believes	believes	4	2	1	3
depends	depends	1	1	1	0
differ	differ	4	2	1	3
difference	difference	1	1	1	0
exposures	exposures	1	1	1	0
nearly	nearly	1	1	1	0
revised	revised	2	2	1	1
risks	risks	6	2	1	5
unspecified	unspecified	2	2	1	1
varies	varies	1	1	1	0

Figure 15- RapidMiner: Uncertainty Words analysis for Continuous Four Quarters



Word	Attribute Na...	Total O...	Docum...	TGT_17Q3	SHLD_17Q3	TGT_17Q4	SHLD_17Q4	TGT_18Q1	SHLD_18Q1	TGT_18Q2	SHLD_18Q2
approximately	approximately	375	8	6	74	21	117	6	65	6	80
intangible	intangible	107	7	0	16	8	43	1	17	1	21
risk	risk	104	8	9	10	26	23	10	8	10	8
assumptions	assumptions	83	8	2	7	14	38	2	7	2	11
believe	believe	67	8	8	6	23	9	5	5	5	6
risks	risks	59	8	1	5	15	24	1	5	1	7
indefinite	indefinite	52	5	0	9	2	20	0	8	0	13
probable	probable	36	5	0	7	4	9	0	7	0	9
independent	independent	33	5	0	2	13	8	0	4	0	6
differences	differences	31	8	2	2	6	9	3	3	3	3
contingencies	contingencies	29	5	0	5	8	6	0	5	0	5
variable	variable	29	7	0	5	1	5	2	7	2	7
exposure	exposure	28	8	2	2	11	5	2	2	2	2
differ	differ	26	8	1	3	3	9	1	4	1	4
anticipated	anticipated	25	8	1	3	3	10	1	4	1	2
believes	believes	22	8	1	3	1	7	1	3	1	5
dependent	dependent	17	5	0	1	10	4	0	1	0	1
anticipate	anticipate	15	8	1	1	4	5	1	1	1	1
predict	predict	15	5	0	2	1	8	0	2	0	2
vary	vary	15	8	1	2	3	5	1	1	1	1
spending	spending	14	5	0	1	1	6	0	3	0	3

(3.2.2): Complexity Words Analysis to Identify Readability using Loughran-McDonald Sentiment Word List (2020)

I conducted another type of readability analysis by applying a complexity word list to measure firm-level complexity, which would be a key measure for a company's financial performance. In the most recent research, "Measuring Firm Complexity," published by Tim Loughran and Bill McDonald, they argued a lack of multidimensional firm-level complexity measurement concepts. They took a novel text-based approach measuring firm-level complexity. Loughran & McDonald (2020) created a list of 374 words that proxy for the complexity of various firm attributes produced by examining actual word usage in U.S. annual reports [TABLE 1]. I applied this complexity word list in RapidMiner and performed text processing as the uncertainty word analysis steps to identify content readability. The results show that the top five complexity words, “lease, subsidiaries, lien, collateral, leaseback,” were more often used in Sears' financial releases than in Target's as anticipated [Figure 16]. In figure 16, the results align with the original

anticipation that complexity words were used more often in Sears' financial releases than in Target's one-year-ahead before Sears bankruptcy.

Figure 16- RapidMiner: Complexity Words analysis for Continuous Four Quarters

Data

Word	Attribute Na...	Tota...	Docum...	TGT_17Q3	SHLD_17Q3	TGT_17Q4	SHLD_17Q4	TGT_18Q1	SHLD_18Q1	TGT_18Q2	SHLD_18Q2
lease	lease	486	8	24	52	60	83	77	52	75	63
subsidiaries	subsidiaries	341	7	0	69	12	106	1	69	1	83
lien	lien	290	4	0	57	0	103	0	59	0	71
collateral	collateral	216	4	0	36	0	78	0	51	0	51
leases	leases	208	8	15	16	30	31	44	16	39	17
leaseback	leaseback	201	4	0	57	0	52	0	43	0	49
entities	entities	176	6	7	25	13	45	0	40	0	46
segment	segment	126	8	12	28	21	18	3	15	1	28
affiliated	affiliated	120	8	1	18	1	33	1	29	1	36
contracts	contracts	114	8	2	13	10	19	3	31	3	33
intangible	intangible	107	7	0	16	8	43	1	17	1	21
affiliates	affiliates	99	5	0	18	1	36	0	20	0	24
foreign	foreign	95	7	0	8	15	36	3	14	3	16
covenants	covenants	94	8	3	15	5	24	4	17	4	22
accrued	accrued	87	8	3	12	12	20	7	13	7	13
subsidiary	subsidiary	62	7	0	14	2	18	1	13	1	13
segments	segments	46	4	0	12	0	11	0	11	0	12
leased	leased	45	8	1	3	7	13	6	4	6	5
litigation	litigation	42	8	3	3	12	14	2	3	2	3
liens	liens	39	4	0	6	0	13	0	8	0	12
royalty	royalty	39	4	0	5	0	17	0	7	0	10
unrecognized	unrecognized	39	5	0	7	8	11	0	6	0	7
warranties	warranties	38	4	0	7	0	10	0	10	0	11
contract	contract	36	8	2	3	4	6	1	9	1	10
global	global	36	8	8	1	6	11	2	3	2	3

(3.3): VADER[15] Sentiment Analysis

Finally, I performed VADER sentiment analysis in RapidMiner to determine the tone (positive versus negative) of the 10-Qs/ 10-Ks of Sears and Target. Sentiment Analysis combines with Natural Language Processing (NLP), text analysis, and computational analysis techniques. It is used to extract and analyze the textual format data to identify the reviewed information's sentiment (Devipriya B, Kalpana Y, 2019). In the sentiment analysis, I used "Management's Discussion and Analysis of financial conditions and results of operations" (MD&A) as the only source which contains the essential forward-looking statements about the companies (Wang et al., 2013, p. 3). The data period range was selected from the year 2017 quarter one to the year 2018

[15] VADER in RapidMiner: using the polarity_score calculator of SentimentIntensityAnalyzer model. This model generates four scores: (i) Negativity, (ii) Positivity, (iii) Neutrality score of the sentence, and finally, (iv) Compound sentiment score of the sentence. The compound score is basically an aggregated version of the first three scores, and will be using this score to measure the sentiment of this study.

quarter two for sentiment analysis since we want to dig more insights into how the companies viewed their near-term objectives for 2017 Q3 onward. I applied ***"Extract Sentiment Operator"*** in RapidMiner and choose "***VADER***" as the execution model [Figure 17]. The generated polarity confidence scores are shown in the following figures [Figure 18].

(3.3.1) How does VADER works on Sentiment Analysis

Sentiment analysis is the process of statistically identifying whether a piece of text is positive, negative, or neutral (Algorithmia blog, 2018). There are two majority approaches ***1) Polarity-based***, where textual data are classified as either positive or negative; or ***2) Valence-based***, where the intensity of the sentiment is considered and calculated. For example, the words "good" and "perfect" would be both classified to positive polarity under polarity-based approach, whereas "perfect" would be designated higher score than "good" under valence-based approach *("Using VADER to Handle Sentiment Analys,"* n.d. 2017)**.** VADER in RapidMiner is a text sentiment analysis model which is sensitive to both polarity (positive/negative) and intensity (strength) of emotion. VADER sentimental analysis relies on VADER's lexicon dictionary that maps lexical features to emotion intensities known as sentiment scores *(Beri, 2020)*. In the VADER model, each of the lexicon words will be rated as positive or negative (polarity-based) and be calculated how the strength of the text's emotion (valence-based). For each piece of text analyzed, sentiment scores measure the strength of the tone from a +1 (strongest positive tone) to -1 (strongest negative tone) via scoring string [Figure 19&20] and add up all the identical scoring string results to come up with a comprehensive polarity score in column "Score" under Data summary. The higher the total score, the more positive the contents [Figure 18].

(3.3.2) VADER Sentiment Analysis results in RapidMiner

In Figure 19, we could notice both Target (blue shade) and Sears (green shade) share a high level of positive polarity scores in the successive six quarters (from 17Q1 to 18Q2); whereas

in Figure 20, Target (blue shade) rated much lower negative polarity scores than Sears (green shade). After adding up positivity score and negativity score, the total score in Figure 18 shows that Target earns a much higher total score than Sears in each quarter analyzed. There could include other various outside factors that could cause the variance in sentiment. Generally, a higher positive sentiment would indicate a more positive outlook on performance. In Figure 21, the sentiment score heatmap provides clear visualization of the total score, negativity score, and positivity score. Target shares a low negativity score (0-1.2 with blue color), while Sears shares a much higher negativity score (2.4 - 4.7 with orange color). However, based on some industry market research reports, the department stores industry has continued its long-term decline over the five years due to the rising e-commerce competition. This finding would be one potential reason Target shares a very similar Sears score in 18Q1 and 18Q2 [Figure 21]. The year 2018 was overall underperforming compared to the year 2017 in this industry. In Figure 22, the broader range box plot of the total score for Target could explain the rising competition in this industry [Figure 22]. Furthermore, sentiment score is not a definite metric for the performance of a company. It is a general indicator that gives some idea of an organization's outlook, so slight differences in sentiment score are likely negligible.

The results of VADER sentiment analysis provide not only remarkable but also encouraging insights. Although the VADER lexicon performs exceptionally well in the social media domain (Hutto & Gilbert, n.d.), the results of this empirical analysis show the advantages could be attained by utilizing VADER in cases of annual reports. Wherein the text data could be a more complex mixture of a range of text.

Figure 17- RapidMiner: Extract Sentiment Operator

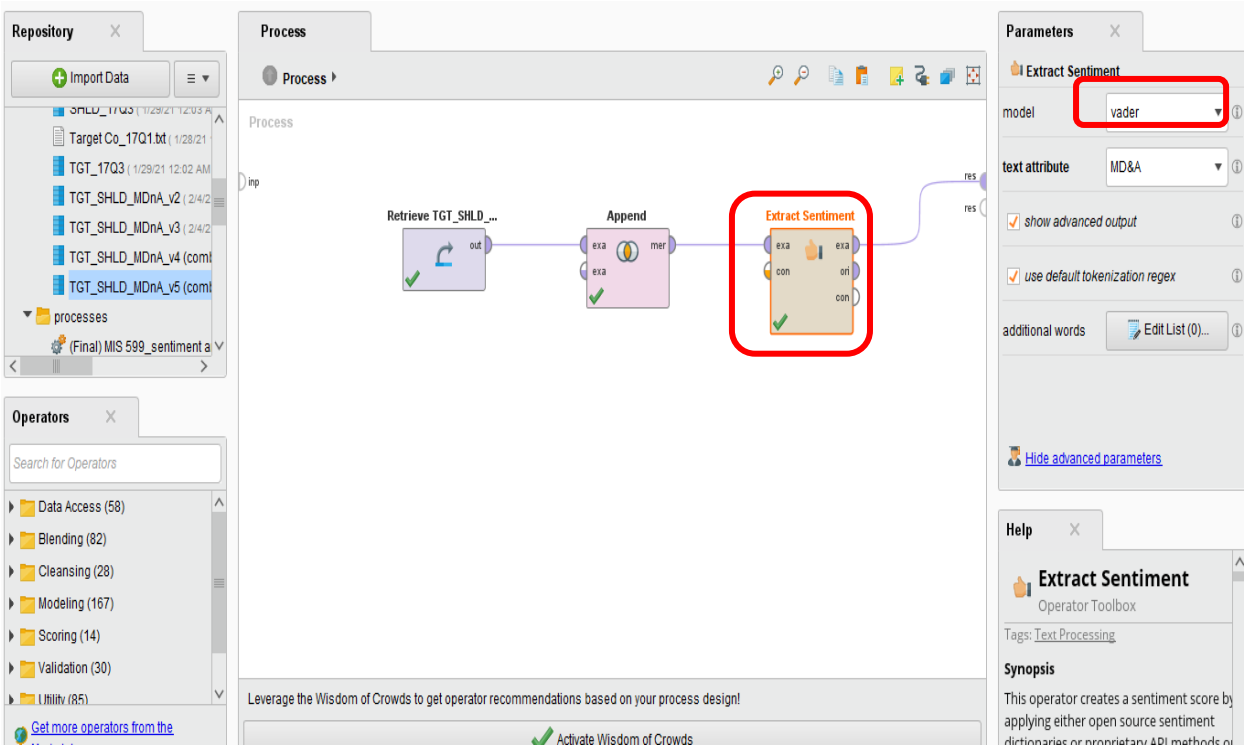
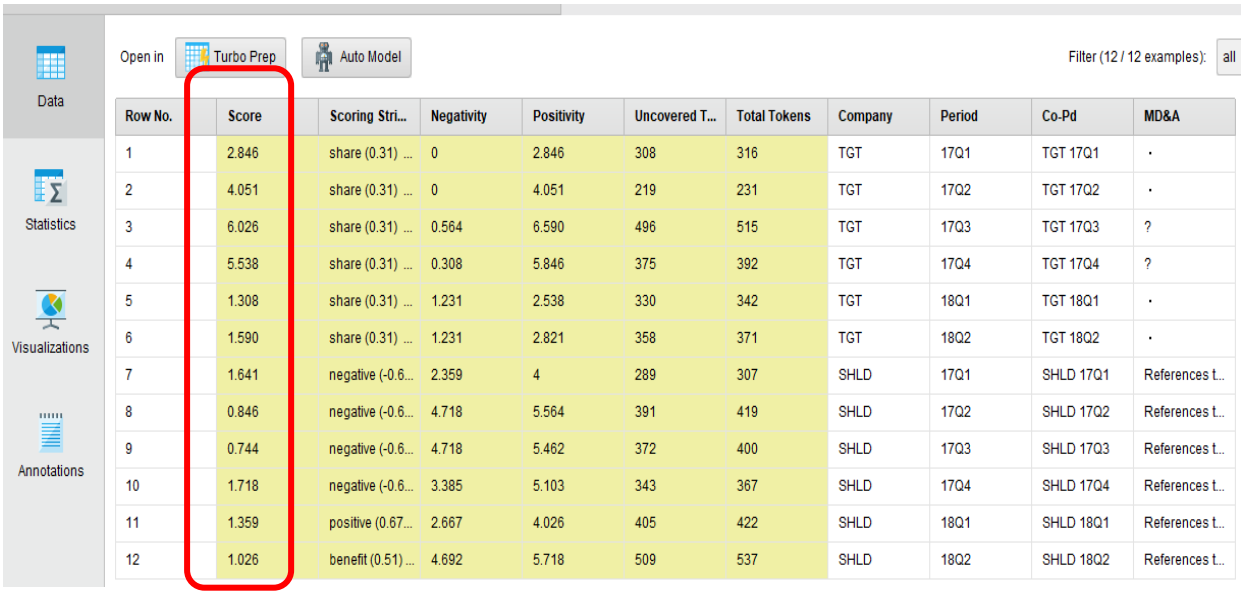


Figure 18- RapidMiner: VADER Sentiment Analysis for Continuous Four Quarters



Open in Turbo Prep | Auto Model | Filter (12 / 12 examples): all

Data | Statistics | Visualizations | Annotations

Row No.	Score	Scoring Stri...	Negativity	Positivity	Uncovered T...	Total Tokens	Company	Period	Co-Pd	MD&A
1	2.846	share (0.31) ...	0	2.846	308	316	TGT	17Q1	TGT 17Q1	·
2	4.051	share (0.31) ...	0	4.051	219	231	TGT	17Q2	TGT 17Q2	·
3	6.026	share (0.31) ...	0.564	6.590	496	515	TGT	17Q3	TGT 17Q3	?
4	5.538	share (0.31) ...	0.308	5.846	375	392	TGT	17Q4	TGT 17Q4	?
5	1.308	share (0.31) ...	1.231	2.538	330	342	TGT	18Q1	TGT 18Q1	·
6	1.590	share (0.31) ...	1.231	2.821	358	371	TGT	18Q2	TGT 18Q2	·
7	1.641	negative (-0.6...	2.359	4	289	307	SHLD	17Q1	SHLD 17Q1	References t...
8	0.846	negative (-0.6...	4.718	5.564	391	419	SHLD	17Q2	SHLD 17Q2	References t...
9	0.744	negative (-0.6...	4.718	5.462	372	400	SHLD	17Q3	SHLD 17Q3	References t...
10	1.718	negative (-0.6...	3.385	5.103	343	367	SHLD	17Q4	SHLD 17Q4	References t...
11	1.359	positive (0.67...	2.667	4.026	405	422	SHLD	18Q1	SHLD 18Q1	References t...
12	1.026	benefit (0.51) ...	4.692	5.718	509	537	SHLD	18Q2	SHLD 18Q2	References t...

Figure 19- RapidMiner: Scoring String (Positivity)

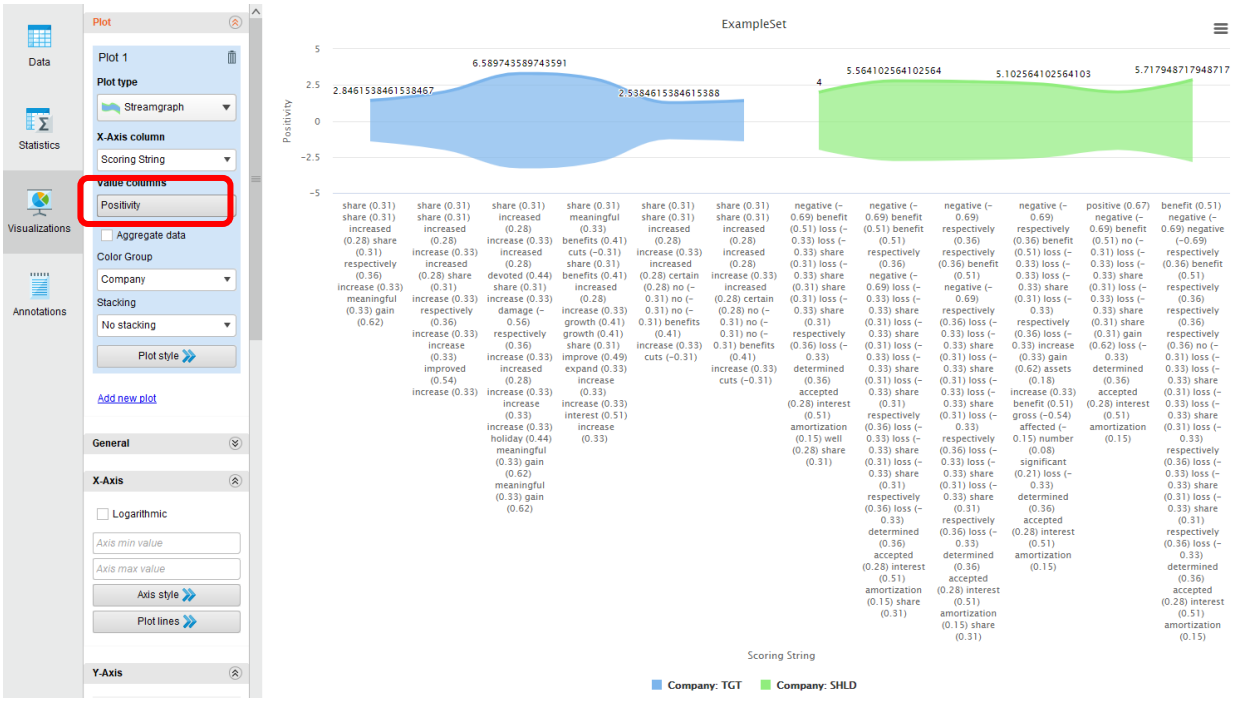


Figure 20- RapidMiner: Scoring String (Negativity)

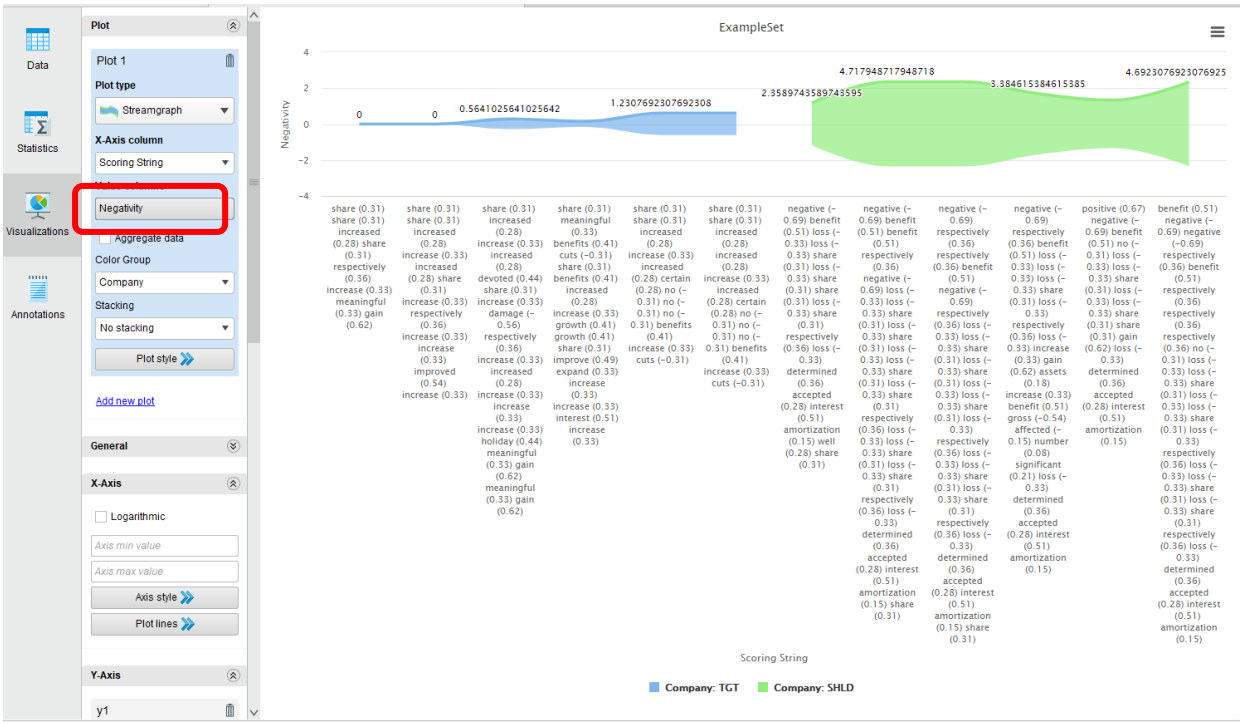


Figure 21- RapidMiner: Scoring Heatmap (Total Score/ Positivity/ Negativity)

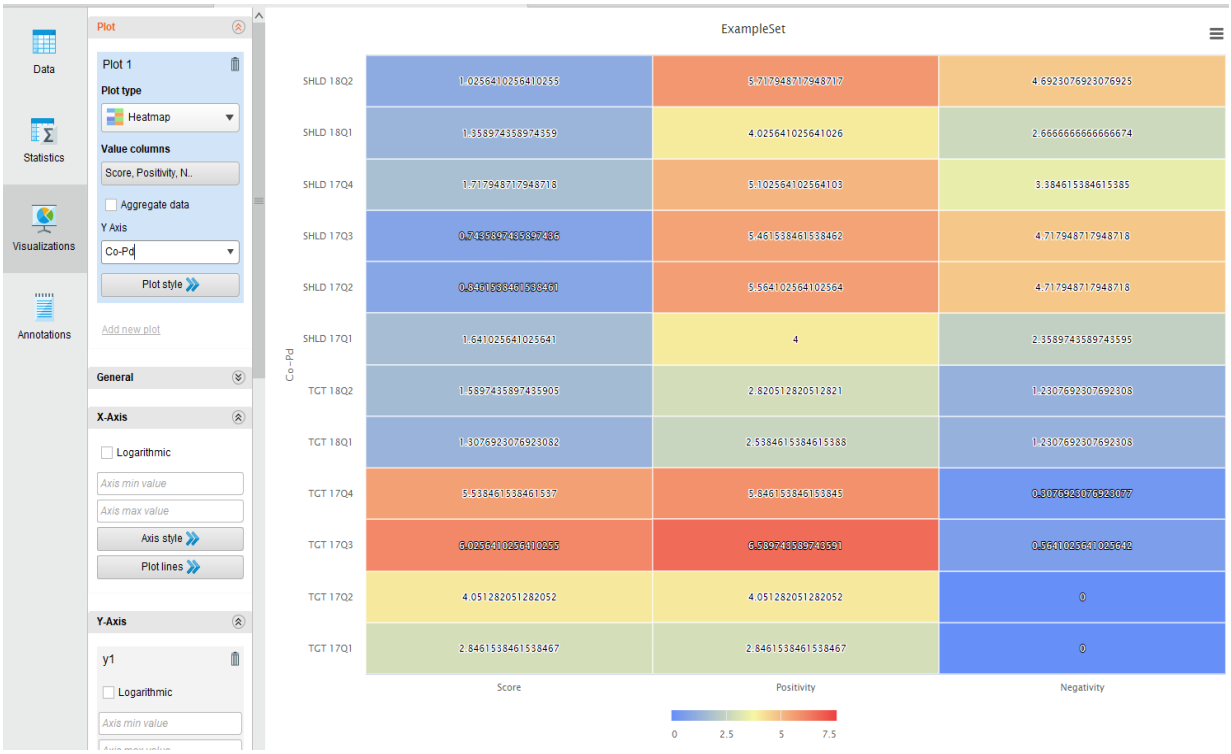
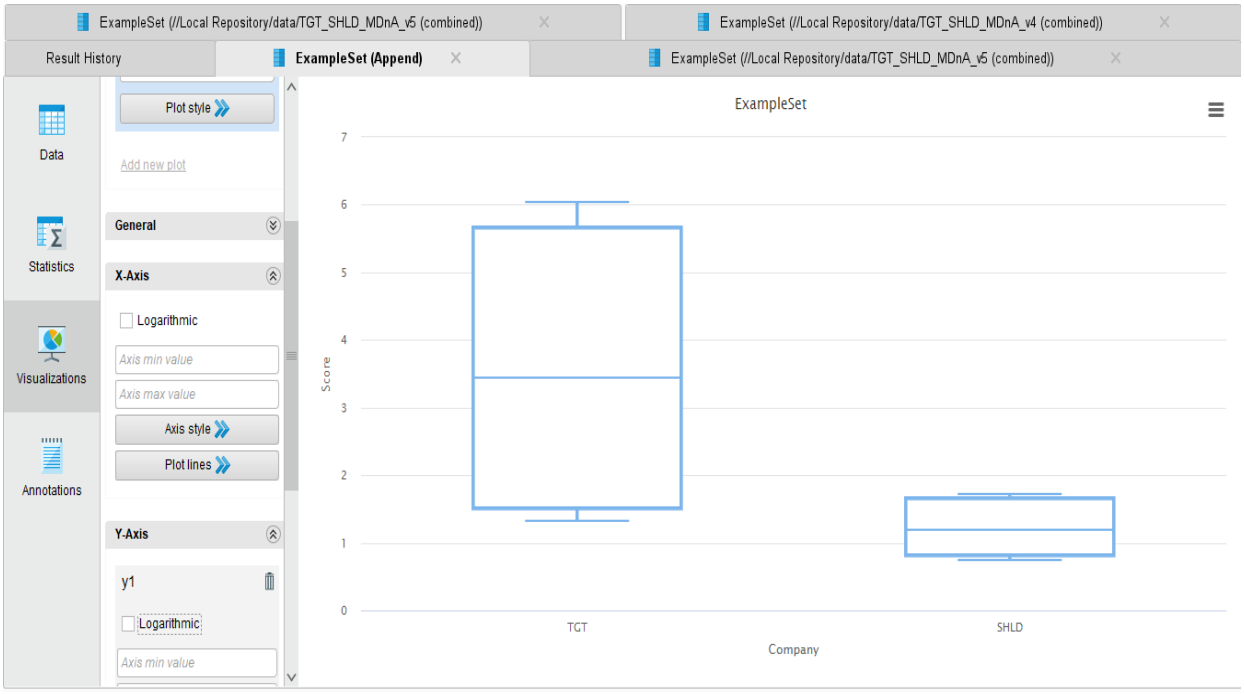


Figure 22- RapidMiner: Score range



(3.4): Evaluate Overall Predictive Analysis for One-Year-Ahead Performance

After performing the various analyses on the 10-Qs and 10-Ks of Target and Sears, it is evident that Sears has a higher risk of not meeting the strong performance out of the two companies. In October 2018, Sears filed for bankruptcy. On the other hand, Target has been doing well and has grown significantly throughout 2019. The analysis also found that Sears has more passive verbs in its 10-Qs and 10-Ks financial statements, suggesting it is trying to avoid ownership for the disclosures in its financial statements. Sears also had more uncertainty and complexity words in its financial statements, indicating a higher risk of fraud and a higher possibility of not achieving its goal and expected performance. After analyzing Sears and Target using VADER, the sentiment results also support the conclusion that Sears is at higher risk of fraud since it generated higher negativity sentiment scores [Figure 20]. Meanwhile, the narrow score range within lower total sentiment scores below 2 [Figure 22] is also a piece of evidence. Overall, the RapidMiner analyses showed that Sears would be more susceptible to its financial performance than Target in one-year-ahead starting from quarter three of 2017.

4.2. AI-based Forecasting Modeling (Quantitative analysis)

This paper also aims to get deeper insights into applying AI techniques to predict intricate sale patterns and compare the results with traditional forecasting models. The retail sales data set[16] was used in this study for applying ***JMP*** statistic modeling both in advanced forecasting method ***ARIMA (Auto-Regressive Integrated Moving Average) and AI-based models*** for comparison. ARIMA is a model given time series based on past sales values. Any non-seasonal time series that exhibits patterns can be modeled with ARIMA models. In JMP, there also offer seasonal ARIMA as a smoothing method for datasets featuring both trends and seasonality.

[16] Retail sales data set: https://www.kaggle.com/aremoto/retail-sales-forecast?select=stores+data-set.csv

However, AI-powered models also consider inputs of influential/causal factors and revisit them every time to identify recent trends related to those factors. Uber's Financial Intelligence-Financial Cloud Platforms ***(section 3.2.2)*** is an example of an AI-powered forecasting model. Information such as seasonality or sentiment statements derived from textual analytics could provide a more robust and precise forecast.

4.2.1. Real Case Implementation via JMP Analysis for Retail sales Forecast

(1) Data Collection

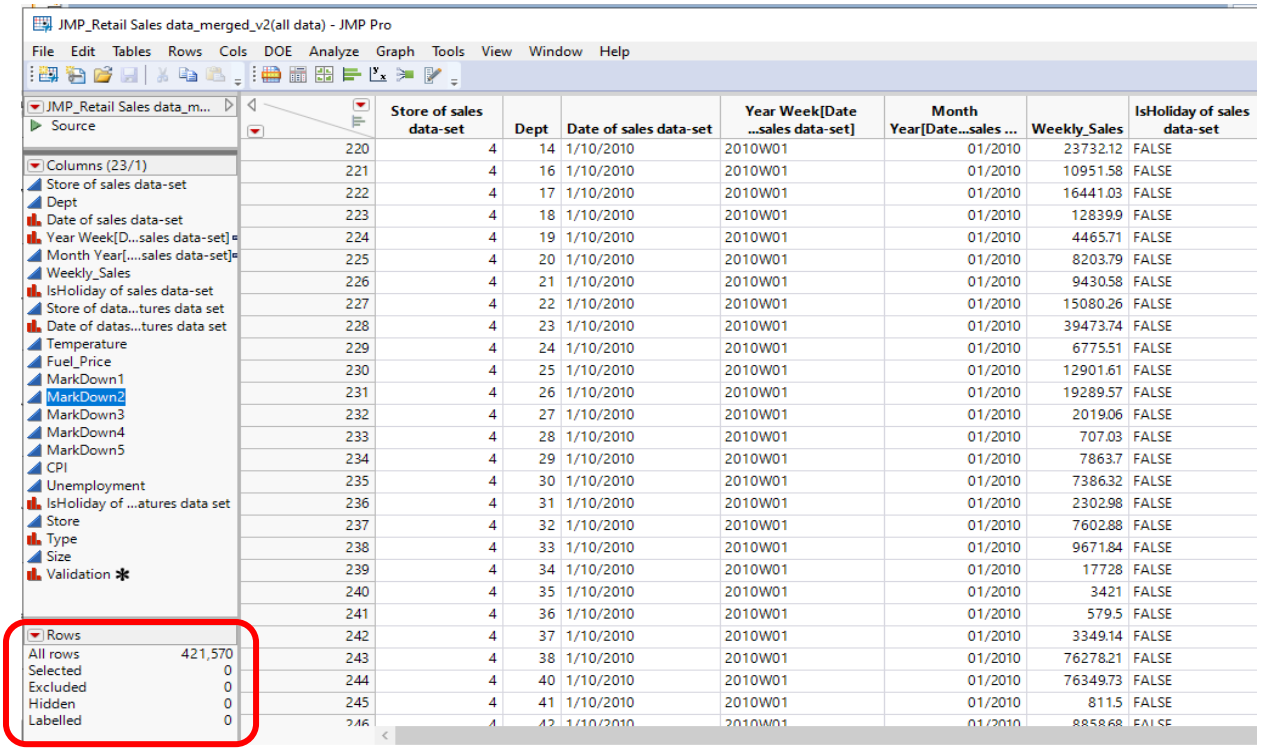
To develop an effective forecasting model, I collected the retail sales dataset with 421,570-row items (Stores*Depts*52 weeks*3years) and 11 attributes (Stores, Depts, Date of sales, IsHoliday of sales, Temperature, Fuel_Price, MarkDown, CPI, Unployment, Type, Size…etc.), which was merged from three identical CSV files in JMP via *JOIN function* and conducted detail analysis after merging the files into one comprehensive file [Figure 23]. The original data sources include the following three tables:

- **Sales dataset:** Historical sales data, which covers periods from 2010-02-05 to 2012-11-01, including the following fields: **Store**[17]**, Dept, Date, Weekly_Sales, IsHoliday**
- **Stores dataset:** Store, Type, Size
- **Features dataset:** Store, Date, Temperature, Fuel_Price, MarkDown1, MarkDown2, MarkDown3, MarkDown4, MarkDown5, CPI, Unemployment, IsHoliday

To merge these three tables, I opened these three files in JMP (all files have to be open to process) and then used the "JOIN" function under the "Tables" bar via matching foreign key attributes (Store and Date). The goal is to create a relationship between these tables [Figure 24]. Once completing the file integration, I looked over the consolidated table near the left-down corner in JMP to check if the number of row items (421,570 rows) aligning with the original CSV file row items. The goal is to assure the data completeness for further analysis [Figure 23].

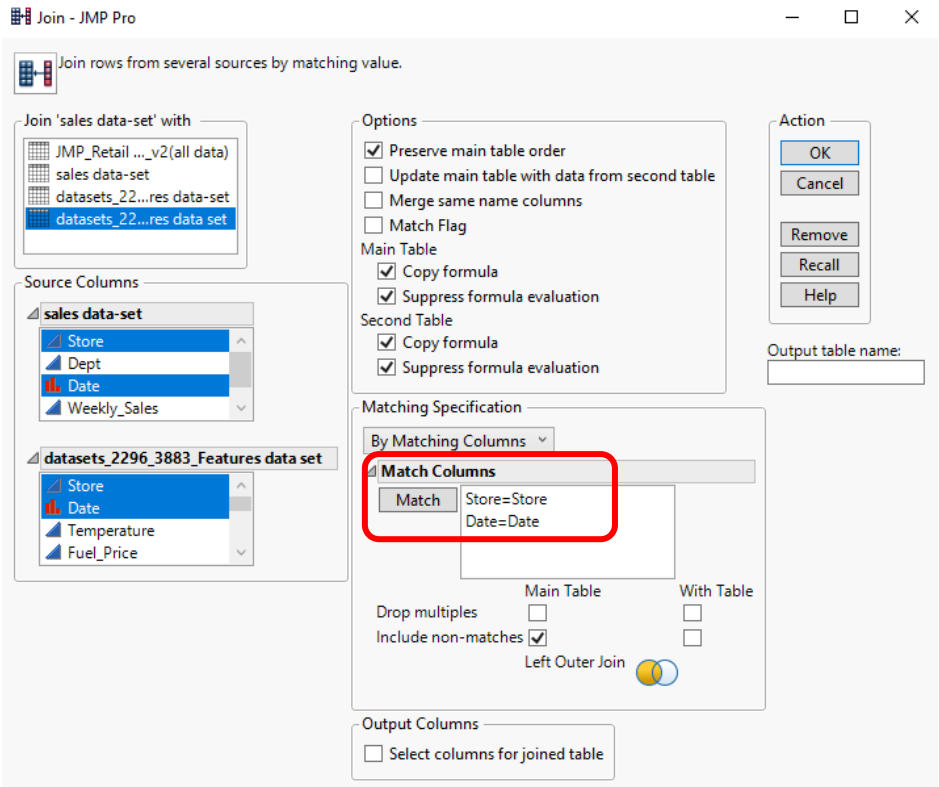
[17] Store and Date are foreign key attributes to merge tables.

Figure 23- JMP: Retail Sales Dataset



	Store of sales data-set	Dept	Date of sales data-set	Year Week[Date ...sales data-set]	Month Year[Date...sales ...	Weekly_Sales	IsHoliday of sales data-set
220	4	14	1/10/2010	2010W01	01/2010	23732.12	FALSE
221	4	16	1/10/2010	2010W01	01/2010	10951.58	FALSE
222	4	17	1/10/2010	2010W01	01/2010	16441.03	FALSE
223	4	18	1/10/2010	2010W01	01/2010	12839.9	FALSE
224	4	19	1/10/2010	2010W01	01/2010	4465.71	FALSE
225	4	20	1/10/2010	2010W01	01/2010	8203.79	FALSE
226	4	21	1/10/2010	2010W01	01/2010	9430.58	FALSE
227	4	22	1/10/2010	2010W01	01/2010	15080.26	FALSE
228	4	23	1/10/2010	2010W01	01/2010	39473.74	FALSE
229	4	24	1/10/2010	2010W01	01/2010	6775.51	FALSE
230	4	25	1/10/2010	2010W01	01/2010	12901.61	FALSE
231	4	26	1/10/2010	2010W01	01/2010	19289.57	FALSE
232	4	27	1/10/2010	2010W01	01/2010	2019.06	FALSE
233	4	28	1/10/2010	2010W01	01/2010	707.03	FALSE
234	4	29	1/10/2010	2010W01	01/2010	7863.7	FALSE
235	4	30	1/10/2010	2010W01	01/2010	7386.32	FALSE
236	4	31	1/10/2010	2010W01	01/2010	2302.98	FALSE
237	4	32	1/10/2010	2010W01	01/2010	7602.88	FALSE
238	4	33	1/10/2010	2010W01	01/2010	9671.84	FALSE
239	4	34	1/10/2010	2010W01	01/2010	17728	FALSE
240	4	35	1/10/2010	2010W01	01/2010	3421	FALSE
241	4	36	1/10/2010	2010W01	01/2010	579.5	FALSE
242	4	37	1/10/2010	2010W01	01/2010	3349.14	FALSE
243	4	38	1/10/2010	2010W01	01/2010	76278.21	FALSE
244	4	40	1/10/2010	2010W01	01/2010	76349.73	FALSE
245	4	41	1/10/2010	2010W01	01/2010	811.5	FALSE

Figure 24- JMP: Merging Retail Sales Dataset



(2) Empirical analysis

(2.1) Time-Series Analysis- Seasonal ARIMA

After the data extraction, transformation, loading (ETL process), and verifying data completeness and data integrity, I performed "Time-Series" analysis under "Specialized Modeling" under the "Analyze" bar. In JMP, the platform provides no-code forecasting modeling such as ARIMA, Seasonal ARIMA, and smoothing Models [Figure 25]. I selected seasonal ARIMA and Seasonal Exponential Smoothing models for this study, and we can review the forecasting results in [Figure 26] for store 1 as an example. This analysis information provides other insights for verifying sales forecasting accuracy. "Time-Series Forecast" is also available in JMP for analysis, which JMP fitting the optimal model in the system and comes out with forecast results. Example results would be seen in Figures 27 and 28 for stores 1 and 20.

Figure 25- JMP: Time-Series Analysis

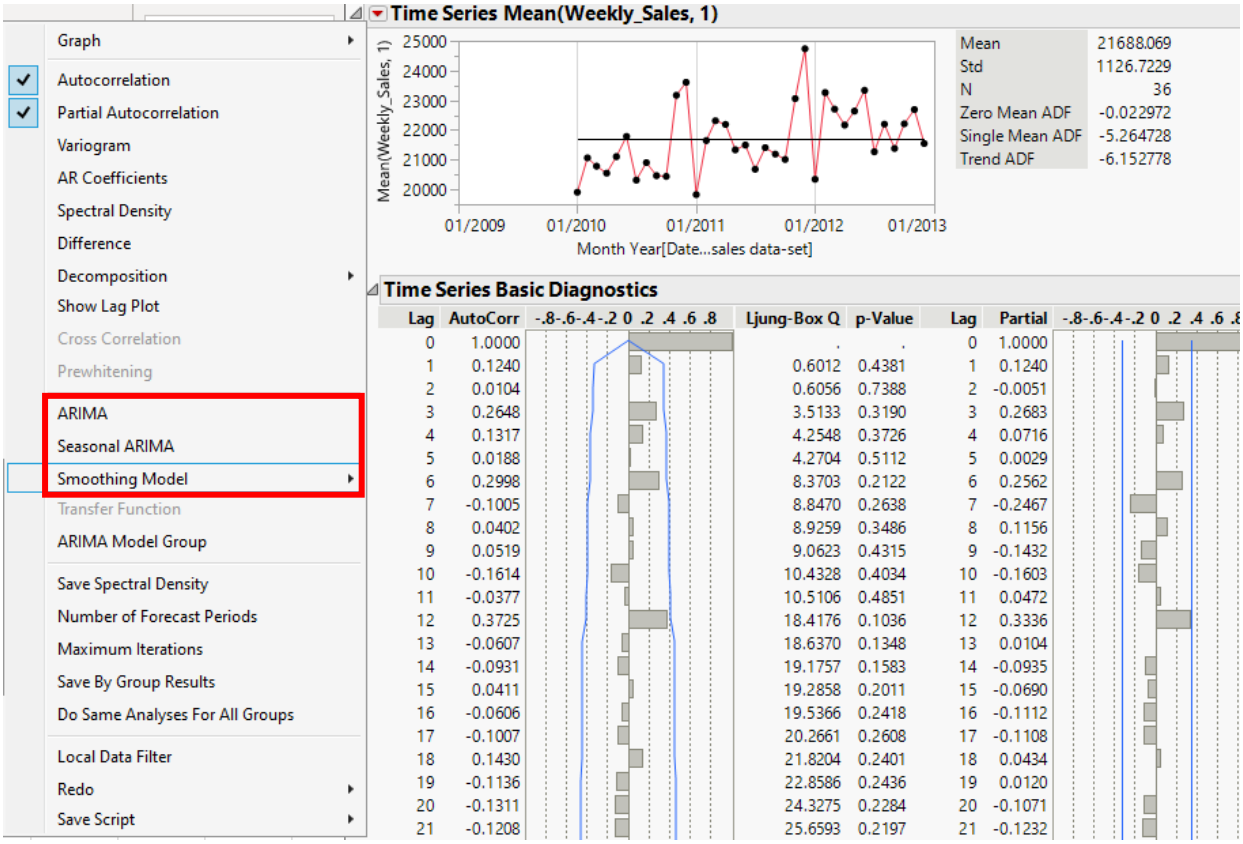


Figure 26- JMP: Seasonal ARIMA / Seasonal Exponential Smoothing model

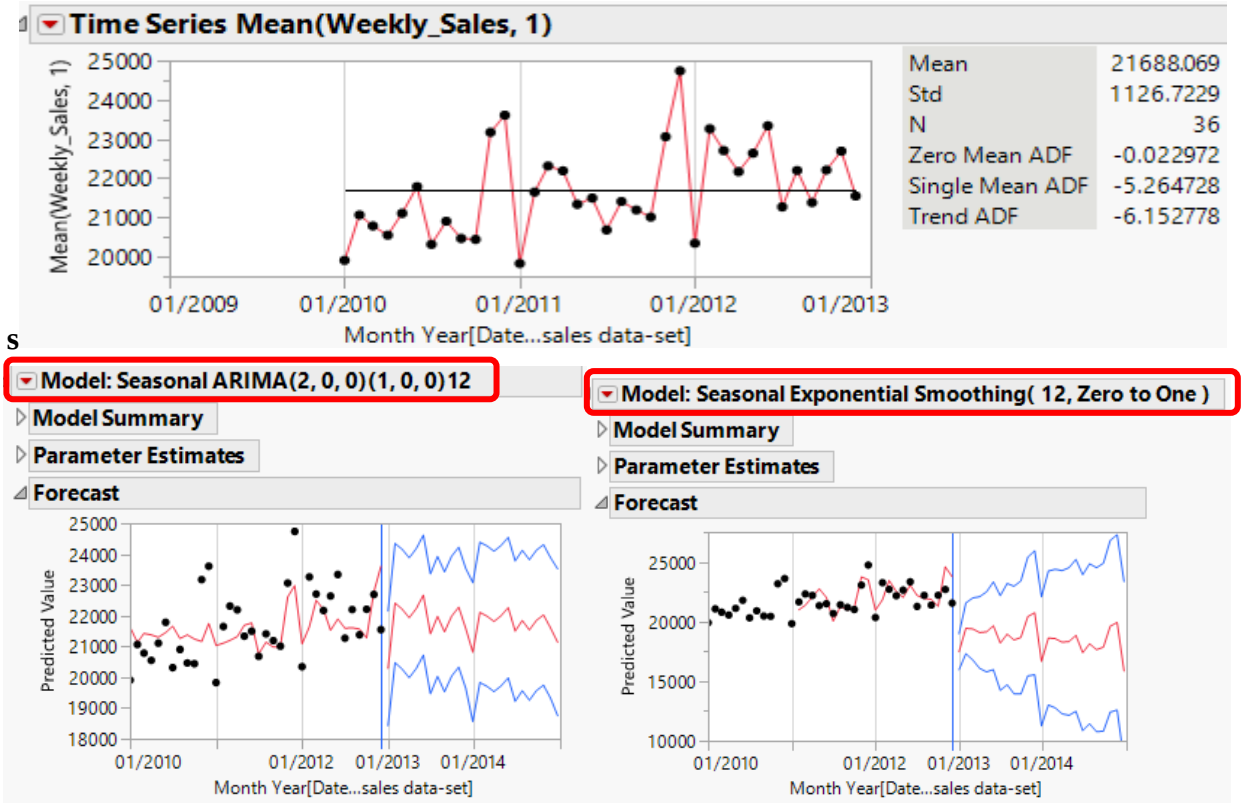


Figure 27- JMP: "Time-Series Forecast" Analysis with Fitted Model (Store 1)

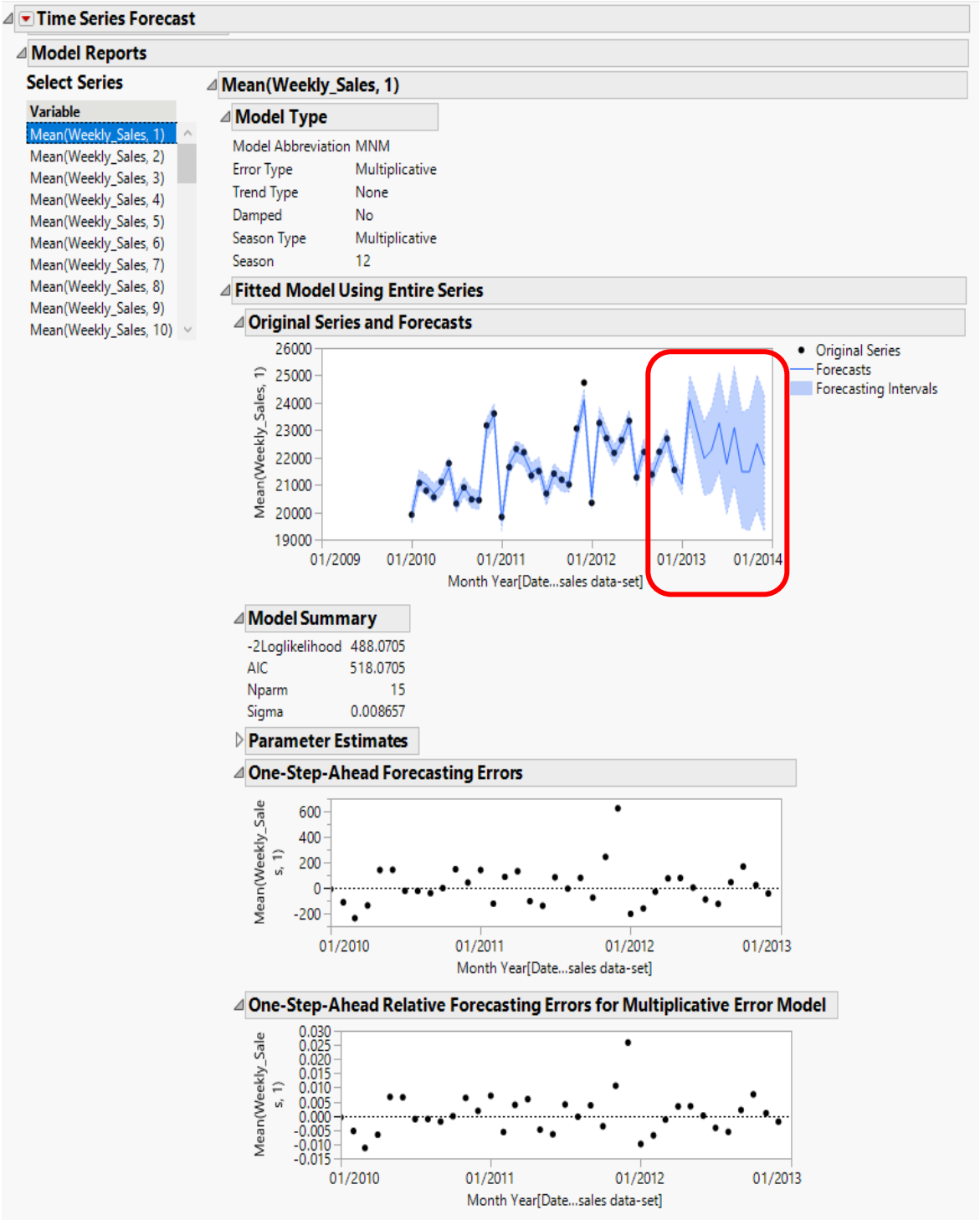
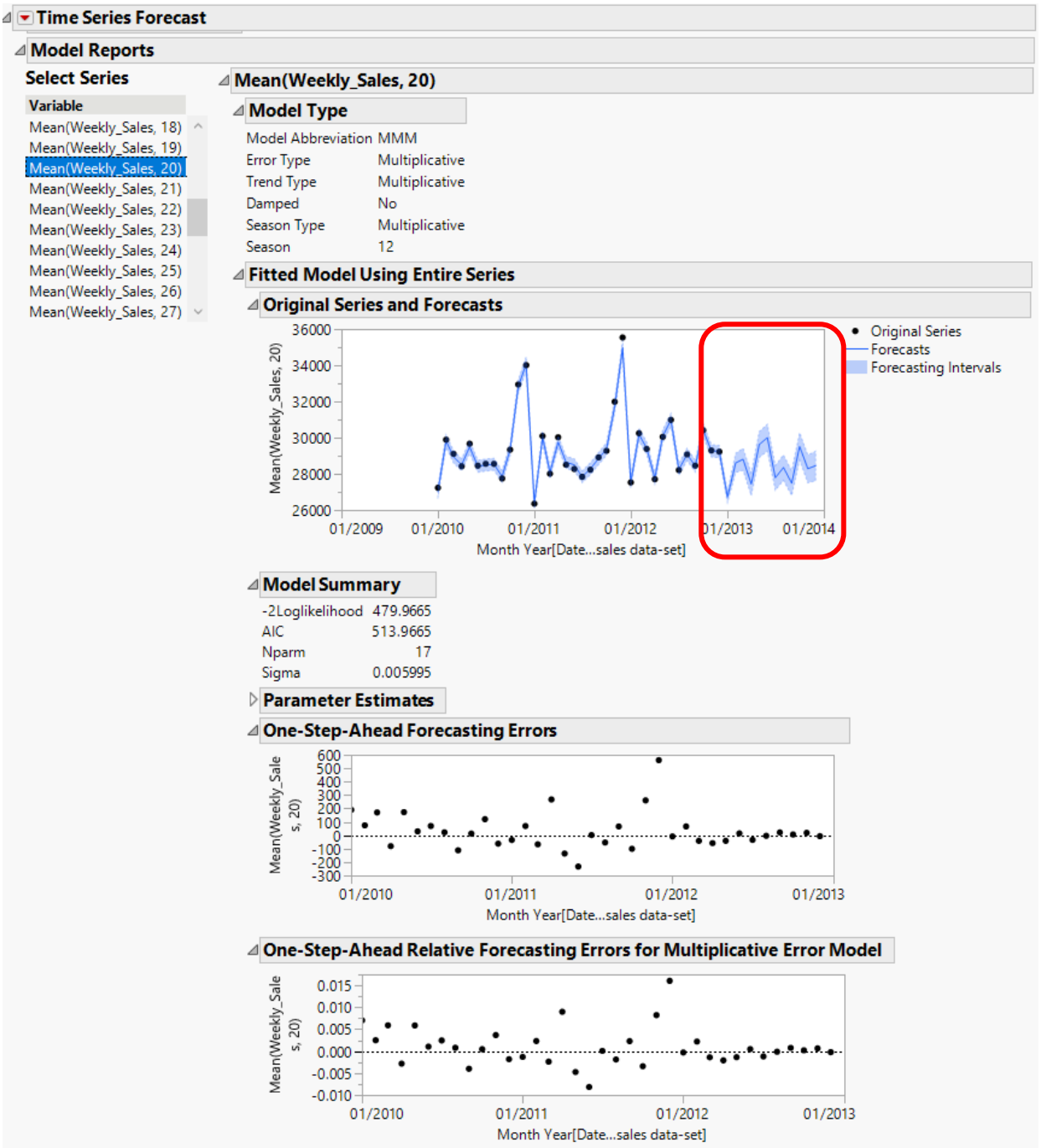


Figure 28- JMP: "Time-Series Forecast" Analysis with Fitted Model (Store 20)

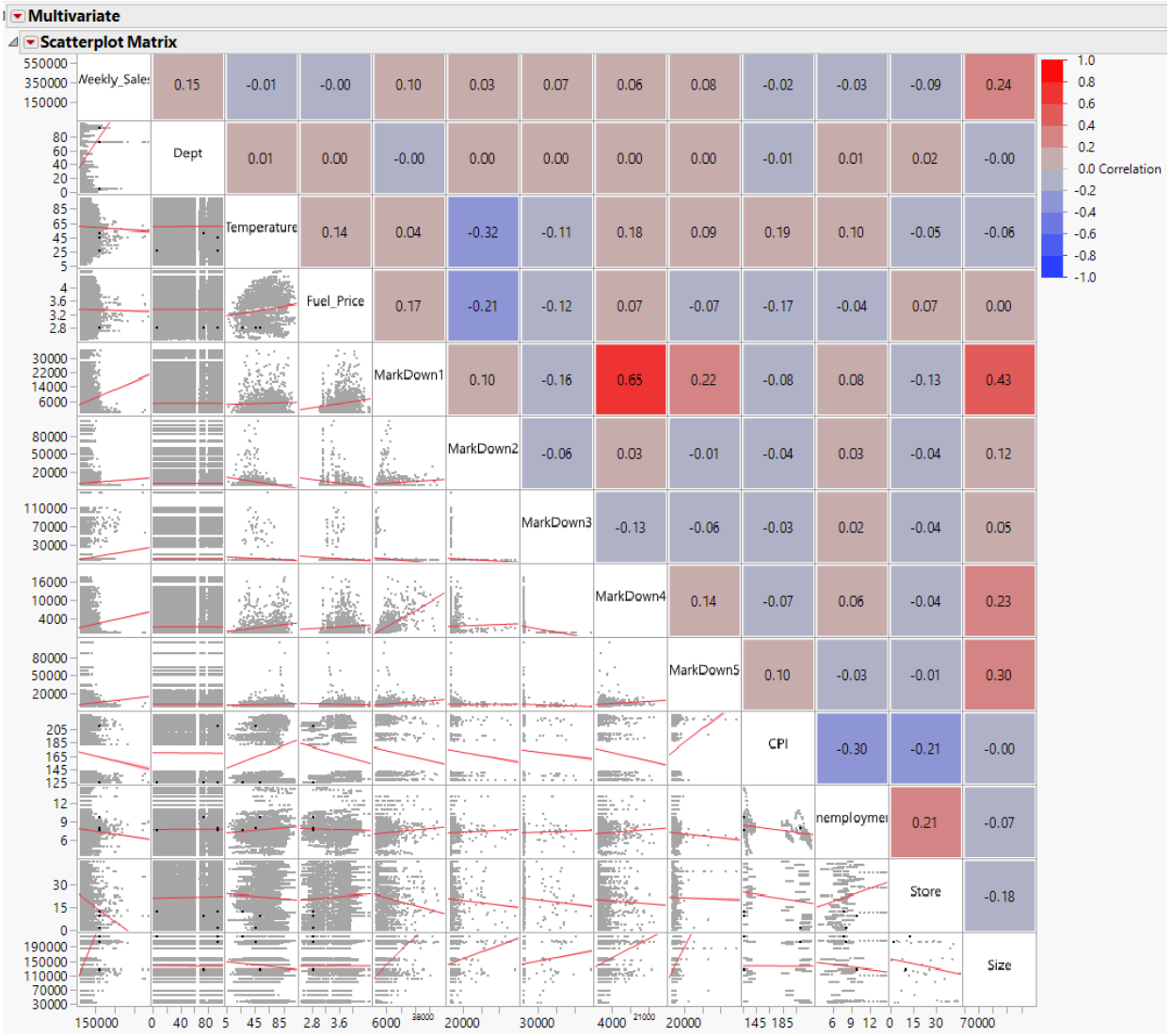


(2.2) AI-based Exploratory Analysis

However, ARIMA is still a traditional time series model, and the forecasting interval [Figure 27 & 28] could cover a wide range, suggesting the inferior model for prediction accuracy. In JMP, I performed **Multivariate Method- Scatterplot Matrix** to gain a brief understanding of the

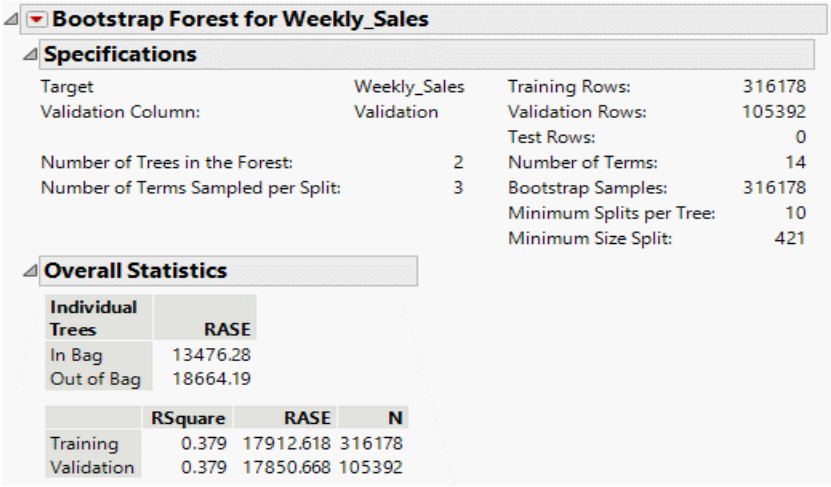
correlation between the response variable (Weekly-Sales) and potential predictor variables [Figure 29]. From the following Scatterplot Matrix, only "Dept" and "Size" have a slightly positive relationship with "Weekly_Sales."

Figure 29- JMP: Multivariate Method- Scatterplot Matrix



To better understand how strong the relationships between the response variable and these potential predictors on this dataset, I ran ***Random Forest*** [Figure 30] and ***Boosted Tree*** [Figure 31]. Both models suggested features "Dept." and "Size" are the top two predictors that substantially explain Weekly_Sales predictions.

Figure 30- JMP: Bootstrap Forest



Bootstrap Forest for Weekly_Sales

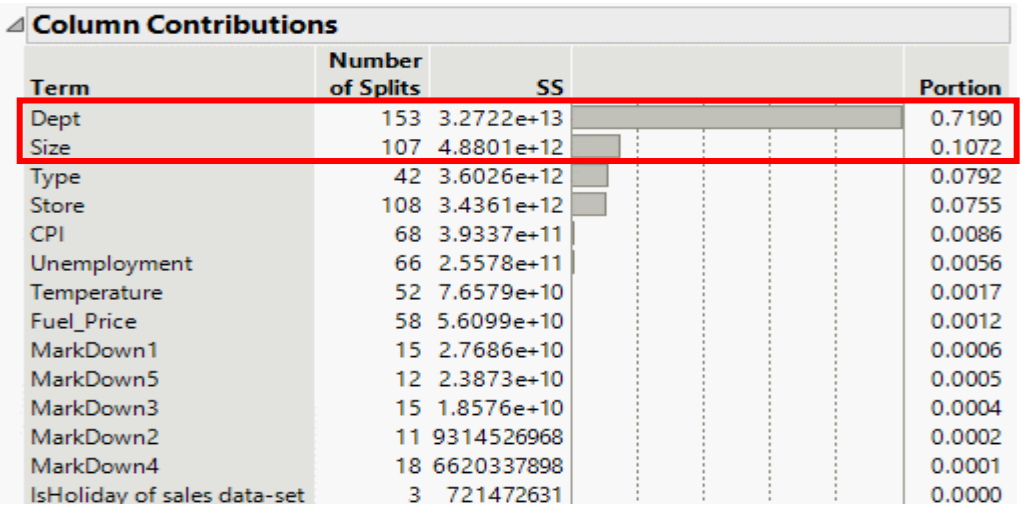
Specifications

Target	Weekly_Sales	Training Rows:	316178
Validation Column:	Validation	Validation Rows:	105392
		Test Rows:	0
Number of Trees in the Forest:	2	Number of Terms:	14
Number of Terms Sampled per Split:	3	Bootstrap Samples:	316178
		Minimum Splits per Tree:	10
		Minimum Size Split:	421

Overall Statistics

Individual Trees	RASE
In Bag	13476.28
Out of Bag	18664.19

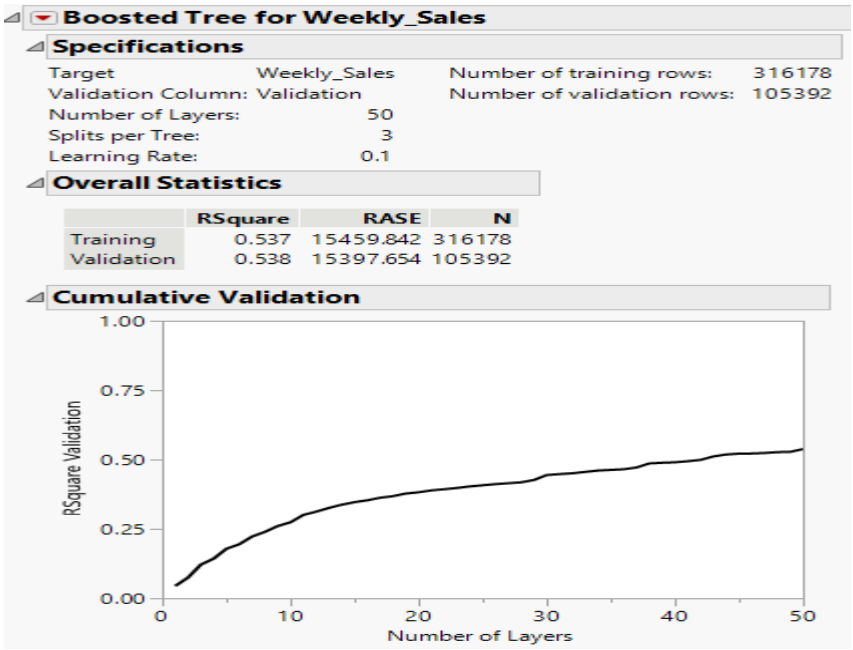
	RSquare	RASE	N
Training	0.379	17912.618	316178
Validation	0.379	17850.668	105392



Column Contributions

Term	Number of Splits	SS	Portion
Dept	153	3.2722e+13	0.7190
Size	107	4.8801e+12	0.1072
Type	42	3.6026e+12	0.0792
Store	108	3.4361e+12	0.0755
CPI	68	3.9337e+11	0.0086
Unemployment	66	2.5578e+11	0.0056
Temperature	52	7.6579e+10	0.0017
Fuel_Price	58	5.6099e+10	0.0012
MarkDown1	15	2.7686e+10	0.0006
MarkDown5	12	2.3873e+10	0.0005
MarkDown3	15	1.8576e+10	0.0004
MarkDown2	11	9314526968	0.0002
MarkDown4	18	6620337898	0.0001
IsHoliday of sales data-set	3	721472631	0.0000

Figure 31- JMP: Boosted Tree



Boosted Tree for Weekly_Sales

Specifications

Target	Weekly_Sales	Number of training rows:	316178
Validation Column:	Validation	Number of validation rows:	105392
Number of Layers:	50		
Splits per Tree:	3		
Learning Rate:	0.1		

Overall Statistics

	RSquare	RASE	N
Training	0.537	15459.842	316178
Validation	0.538	15397.654	105392

Cumulative Validation

Column Contributions

Term	Number of Splits	SS	Portion
Dept	115	3.6905e+14	0.7987
Size	27	8.7353e+13	0.1891
Store	7	5.5587e+12	0.0120
MarkDown3	1	9.4144e+10	0.0002
IsHoliday of sales data-set	0	0	0.0000
Temperature	0	0	0.0000
Fuel_Price	0	0	0.0000
MarkDown1	0	0	0.0000
MarkDown2	0	0	0.0000
MarkDown4	0	0	0.0000
MarkDown5	0	0	0.0000
CPI	0	0	0.0000
Unemployment	0	0	0.0000
Type	0	0	0.0000

At the current stage, AI-based models are primarily used for exploratory analysis only. Kumar et al. (2020) used the time series model as a baseline and applied fuzzy artificial neural networks (fuzzy ANN) based classifier as an AI-based forecasting model. This pioneer foretasting methodology is still in the research stage, while rarely in business practice.

(3) Evaluate AI-Based Forecasting Model for Retail Sales Performance

Machine Learning and big data techniques are bringing paradigm changes in developing a better financial forecasting model. It can analyze vast amounts of data and provide instant insights that can significantly improve business performances. Through the JMP platform, we could identify more insights and information to help corporate finance conducting sales forecasting. Leveraging technology is a way to improve the corporate finance landscapes and accelerate its various offerings to business decision-making.

As Begenau et al. (2018) had mentioned, information technology is pervasive and transformative in modern financial markets. Faster processors crunch ever more data such as macro announcements, earnings statements, and even competitors' performance metrics. All those might provide real-time information to forecast future returns. However, the lack of leveraging qualitative information into forecasting procedures would leave us achieving another milestone in optimizing data technology.

5. CHALLENGES AND FUTURE SCOPE

Though Textual mining and AI-based forecasting would create more accurate, more related real-time information to support business decision planning, previous research also revealed the challenges impeding AI implementation in financial forecasting. Two significant areas pose difficulties in implementing an AI-powered forecasting system (Antwi et al., 2019). One is the high implementation costs, including high-computation-power facilities, data scientists, and technical experts; those are wildly unaffordable for small businesses. The other concern would be the potential for cyber-attacks, primarily through cloud-based enterprise performance management (EPM) or shared-source platforms (Antwi et al., 2019).

This research focuses on reviewing affordable text analytical technology and AI-based forecasting methods that can be used by an internal corporate finance team, rather than outsourcing to technical consultants or internally hiring expensive data scientists. The digital and data revolution provides a massive opportunity for finance professionals. Though they don't have to become expert data scientists, they need higher data science and analytical skills to derive insights from data and deliver foresight to enable more effective decision-making and control. The focus of this research is on upskilling existing finance talents in business organizations.

6. CONCLUSION

For each business, corporate financial forecasting helps in the future prediction of the market trends; it also helps a business evaluate its past and therefore be able to have a more explicit focus on the future. An efficient forecasting process provides quick insights and helps companies plan better, optimize resource allocation efficacy, manage inventory, remain competitive, and keep business sustainable. Thus, text analytics and AI-based forecast methodology must be intelligent, agile, and accurately reflect market dynamics. It should decode relevant signals from the structured and unstructured data gathered from internal and external sources and then factor them

into projection models (Ganpact, 2019). The future of business relies heavily on forecasting, which directly impacts the global economy (Antwi et al., 2019). But, not every company will have the power to own the forecasting technology due to the cost, and businesses will need to increase security to protect the forecasting systems. Annor, Albert et al. (2019) performed identified hypothesis tests to evaluate both the advantage (potential benefits & opportunities) and the disadvantage (potential costs & risks) of the AI-based forecasting model. Businesses can significantly benefit from big-data-driven AI-based forecasting with emerging text analytics technology to explore more business strategic planning precision. However, the "GO" or "No-GO" decision is still pending on high implementation costs in the early stage and cybersecurity loopholes after implementation. The trade-off decision has to be cautiously evaluated. Efficiency and effectiveness are still needed by the price to be reduced (Antwi et al., 2019).

In this paper, we performed qualitative and quantitative analysis, respectively. In the future, we see more potential data analytical technology to leverage qualitative and quantitative research into one comprehensive model to boost forecasting accuracy and accelerate business strategic planning.

ACKNOWLEDGMENTS

Foremost, I would like to express my sincere gratitude toward Dr. Anthony Townsend, my major professor of this course, for his mentorship in the early stage and his continuous support and guidance throughout this graduate research studies. I want to express my sincere appreciation to Dr. Valentina Salotti that her finance professional guiding me with new thoughts and ideas for this project. I would also like to express my deep gratitude toward Dr. Diane Janvrin for her class inspiring me to conduct more profound research on this topic. This project would not have been possible without their precious support and guidance.

Equally importantly, I would like to thank the graduate college faculty and staff members and all the professors for their constant support and guidance throughout my graduate studies. That makes my time at Iowa State University an incredible learning journey.

Additionally, the same appreciation goes to my family, friends, and peers for their unfailing emotional support, timely encouragement, and endless patience and inspiration.

REFERENCES

Alles, M., & Gray, G. (2016). Incorporating big data in audits: Identifying inhibitors and a research agenda to address those inhibitors. *International Journal of Accounting Information Systems*, *22*. https://doi.org/10.1016/j.accinf.2016.07.004

Antwi, A. A., Al-Dherasi, A. A. M., & Chunting, Y. (2019). Application of Artificial Intelligence in Forecasting: A Systematic Review. *International Journal of Computer Applications*, *177*(26), 5–10.

Begenau, J., Farboodi, M., & Veldkamp, L. (2018). Big data in finance and the growth of large firms. *Journal of Monetary Economics*, *97*, 71–87. https://doi.org/10.1016/j.jmoneco.2018.05.013

Beri, A. (2020, May 27). *SENTIMENTAL ANALYSIS USING VADER*. Medium. https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664

Chase, C. W. (2013). *Demand-Driven Forecasting: A Structured Approach to Forecasting*. John Wiley & Sons.

Dame, M. C. W. // U. of N. (n.d.-a). *Complexity // Software Repository for Accounting and Finance // University of Notre Dame*. Software Repository for Accounting and Finance. Retrieved January 29, 2021, from https://sraf.nd.edu/data/complexity/

Dame, M. C. W. // U. of N. (n.d.-b). *Resources // Software Repository for Accounting and Finance // University of Notre Dame*. Software Repository for Accounting and Finance. Retrieved February 5, 2021, from https://sraf.nd.edu/textual-analysis/resources/

Devipriya B, Kalpana Y. (2019). Evaluation of Sentiment Data using Classifier Model in Rapid Miner Tool. (2019). *International Journal of Engineering and Advanced Technology*, *9*(1), 2966–2972. https://doi.org/10.35940/ijeat.A1323.109119

Ertek, D. G., & Ertek, D. G. (2017, October 31). Text Mining with RapidMiner—Dr. Gurdal Ertek's Publications. *Dr. Gürdal Ertek's Publications*. https://ertekprojects.com/gurdal-ertek-publications/text-mining-with-rapidminer/

Gepp, A., Linnenluecke, M., O'Neill, T., & Smith, T. (2018). Big Data Techniques in Auditing Research and Practice: Current Trends and Future Opportunities. *Journal of Accounting Literature*, *40*, 102–115. https://doi.org/10.2139/ssrn.2930767

Ghosh, S. (2019, January 24). Top B2B Technology Companies for AI in Sales Forecasting. *SalesTech Star*. https://salestechstar.com/predictive-analytics/top-b2b-technology-companies-for-ai-in-sales-forecasting/

Guo, L., Shi, F., & Tu, J. (2016). Textual analysis and machine learning: Crack unstructured data in finance and accounting. *The Journal of Finance and Data Science*, *2*(3), 153–170. https://doi.org/10.1016/j.jfds.2017.02.001

Gupta, A., Dengre, V., Kheruwala, H. A., & Shah, M. (2020). Comprehensive review of text-mining applications in finance. *Financial Innovation*, *6*(1), 39. https://doi.org/10.1186/s40854-020-00205-1

Hasan, Md. M., Popp, J., & Oláh, J. (2020). Current landscape and influence of big data on finance. *Journal of Big Data*, *7*(1), 21. https://doi.org/10.1186/s40537-020-00291-z

How Machine Learning Can Transform The Financial Forecasting Process | by Neevista Pty Ltd | Medium. (n.d.). Retrieved January 15, 2021, from https://medium.com/@neevista/how-machine-learning-can-transform-the-financial-forecasting-process-357bfd87c2ba

Hutto, C. J. (2021). *Cjhutto/vaderSentiment* [Python]. https://github.com/cjhutto/vaderSentiment (Original work published 2014)

Hutto, C. J., & Gilbert, E. (n.d.). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. 10.

Hogan, S. C., & Merrill, E. (n.d.). *Algorithmic Forecasting in a Digital World: Crunch Time Series*. Deloitte United States. Retrieved January 28, 2021, from https://www2.deloitte.com/us/en/pages/finance-transformation/articles/algorithmic-analytics-to-improve-forecasting-process.html

Introduction to sentiment analysis: What is sentiment analysis? (2018, March 26). *Algorithmia Blog*. https://algorithmia.com/blog/introduction-sentiment-analysis

Kumar, A., Shankar, R., & Aljohani, N. R. (2020). A big data-driven framework for demand-driven forecasting with effects of marketing-mix variables. *Industrial Marketing Management*, *90*, 493–507. https://doi.org/10.1016/j.indmarman.2019.05.003

Lau, R. Y. K., Zhang, W., & Xu, W. (2018). Parallel Aspect-Oriented Sentiment Analysis for Sales Forecasting with Big Data. *Production and Operations Management*, *27*(10), 1775–1794. https://doi.org/10.1111/poms.12737

Lee, B., Park, J.-H., Kwon, L., Moon, Y.-H., Shin, Y., Kim, G., & Kim, H. (2018). About relationship between business text patterns and financial performance in corporate data. *Journal of Open Innovation: Technology, Market, and Complexity*, *4*. https://doi.org/10.1186/s40852-018-0080-9

Lewis, C., & Young, S. (2019). Fad or future? Automated analysis of financial text and its implications for corporate reporting. *Accounting and Business Research*, *49*(5), 587–615. https://doi.org/10.1080/00014788.2019.1611730

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, *45*(2–3), 221–247. https://doi.org/10.1016/j.jacceco.2008.02.003

Liu, M., Wu, K., Yang, R., & Yu, Y. (2020, July 15). Textual Analysis for Risk Profiles from 10-K Filings. *The CPA Journal*. https://www.cpajournal.com/2020/07/15/textual-analysis-for-risk-profiles-from-10-k-filings/

Lolli, F., Gamberini, R., Regattieri, A., Balugani, E., Gatos, T., & Gucci, S. (2017). Single-hidden layer neural networks for forecasting intermittent demand. *International Journal of Production Economics*, *183*, 116–128. https://doi.org/10.1016/j.ijpe.2016.10.021

Loughran, T., & Mcdonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, *66*(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Loughran, T., & Mcdonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research*, *54*(4), 1187–1230. https://doi.org/10.1111/1475-679X.12123

Loughran, T., & McDonald, B. (2020). *Measuring Firm Complexity* (SSRN Scholarly Paper ID 3645372). Social Science Research Network. https://doi.org/10.2139/ssrn.3645372

Lucas, S. (2012, August 21). *Beyond the Balance Sheet: Run Your Business on New Signals in the Age of Big Data*. SAP HANA. https://blogs.saphana.com/2012/08/21/beyond-the-balance-sheet-run-your-business-on-new-signals-in-the-age-of-big-data/

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLOS ONE*, *13*(3), e0194889. https://doi.org/10.1371/journal.pone.0194889

Neevista Pty Ltd. (2020, January 20). *Predicting SL/TP Signal Using Machine Learning*. QuantInsti. https://blog.quantinsti.com/predicting-sl-tp-signal-machine-learning-project-sunanda/

Nguyen, B.-H., & Huynh, V.-N. (2020). Textual analysis and corporate bankruptcy: A financial dictionary-based sentiment approach. *Journal of the Operational Research Society*, *0*(0), 1–20. https://doi.org/10.1080/01605682.2020.1784049

Pejić Bach, M., Krstić, Ž., Seljan, S., & Turulja, L. (2019). Text Mining for Big Data Analysis in Financial Sector: A Literature Review. *Sustainability*, *11*(5), 1277. https://doi.org/10.3390/su11051277

Shivang. (2018, September 16). Uber's Financial Cloud Platform—How A Massive Global Ride Sharing Company is Forecasting it's Finances with Machine Learning & Data Science. *8bitmen.Com*. https://www.8bitmen.com/ubers-financial-cloud-platform-how-a-massive-global-ride-sharing-company-is-forecasting-its-finances-with-machine-learning-data-science/

Song, C. (2018, July 5). *Transforming Financial Forecasting with Data Science and Machine Learning at Uber*. Uber Engineering Blog. https://eng.uber.com/transforming-financial-forecasting-machine-learning/

Using VADER to handle sentiment analysis with social media text. (n.d.). Retrieved February 11, 2021, from https://t-redactyl.io/blog/2017/04/using-vader-to-handle-sentiment-analysis-with-social-media-text.html

Wang, C.-J., Tsai, M.-F., Liu, T., & Chang, C.-T. (2013). *Financial Sentiment Analysis for Risk Prediction*.

Xing, F. Z., Cambria, E., & Welsch, R. E. (2018). Natural language based financial forecasting: A survey. *Artificial Intelligence Review*, *50*(1), 49–73. https://doi.org/10.1007/s10462-017-9588-9

Yu, T. H.-K., & Huarng, K.-H. (2010). A neural network-based fuzzy time series model to improve forecasting. *Expert Systems with Applications*, *37*(4), 3366–3372. https://doi.org/10.1016/j.eswa.2009.10.013

RESOURCES

[1] URL to the CC papers in the past:
https://lib.dr.iastate.edu/isba_creativecomponents/

[2] Communication Center of the Ivy College for improving writing:
https://www.ivybusiness.iastate.edu/communications-center/

[3] ISU library:
https://www.lib.iastate.edu/research-tools/research-help/find-articles

[4] RapidMiner:
https://my.rapidminer.com/nexus/account/index.html#downloads

[5] U.S. Securities and Exchange Commission- EDGAR | Company Filings Data Search:
https://www.sec.gov/edgar/searchedgar/companysearch.html

[6] LoughranMcDonald_ComplexityWordLists_2020 (List of 374 Complexity Words):
https://sraf.nd.edu/data/complexity/

TABLE 1 - List of 374 Complexity Words (Loughran & McDonald, 2020, p. 36)
(This table presents the 374 words included in the COMPLEXITY lexicon.)

ACCRUABLE	CONVERTIBILITY	LEASEBACK	PATENTING	SEGMENTING
ACCRUAL	CONVERTIBLE	LEASEBACKS	PATENTS	SEGMENTS
ACCRUALS	CONVERTIBLES	LEASED	REACQUIRE	SOVEREIGN
ACCRUE	COPYRIGHT	LEASEHOLD	REACQUIRED	SOVEREIGNS
ACCRUED	COPYRIGHTABLE	LEASEHOLDER	REACQUIRES	SOVEREIGNTIES
ACCRUES	COPYRIGHTED	LEASEHOLDERS	REACQUIRING	SOVEREIGNTY
ACCRUING	COPYRIGHTING	LEASEHOLDS	REACQUISITION	SUBCONTRACT
ACQUIRE	COPYRIGHTS	LEASER	REACQUISITIONS	SUBCONTRACTED
ACQUIRED	COUNTERPARTIES	LEASES	RECAPITALIZATION	SUBCONTRACTING
ACQUIREE	COUNTERPARTY	LEASING	RECAPITALIZATIONS	SUBCONTRACTOR
ACQUIREES	COVENANT	LESSEE	RECAPITALIZE	SUBCONTRACTORS
ACQUIRER	COVENANTED	LESSEES	RECAPITALIZED	SUBCONTRACTS
ACQUIRERS	COVENANTING	LESSOR	RECAPITALIZES	SUBLEASE
ACQUIRES	COVENANTS	LESSORS	RECAPITALIZING	SUBLEASED
ACQUIRING	DERIVATIVE	LICENCE	RECLASSIFICATION	SUBLEASEE
ACQUIROR	DERIVATIVES	LICENCED	RECLASSIFICATIONS	SUBLEASEHOLD
ACQUIRORS	EMBEDDED	LICENCES	RECLASSIFIED	SUBLEASES
ACQUISITION	ENTITIES	LICENCING	RECLASSIFIES	SUBLEASING
ACQUISITIONS	EXERCISABILITY	LICENSABLE	RECLASSIFY	SUBLESSEE
ACQUISITIVE	EXERCISABLE	LICENSE	RECLASSIFYING	SUBLESSEES
AFFILIATE	EXERCISEABILITY	LICENSED	REISSUANCE	SUBLESSOR
AFFILIATED	EXERCISEABLE	LICENSEE	REISSUANCES	SUBLESSORS
AFFILIATES	EXERCISED	LICENSEES	REISSUE	SUBLET
AFFILIATING	FLOATING	LICENSES	REISSUED	SUBLETS
AFFILIATION	FOREIGN	LICENSING	REISSUES	SUBLETTING
AFFILIATIONS	FRANCHISE	LICENSOR	REISSUING	SUBLETTINGS
ALLIANCE	FRANCHISED	LICENSORS	REORGANISATION	SUBLICENSABLE
ALLIANCES	FRANCHISEE	LIEN	REORGANIZATION	SUBLICENSE
BANKRUPT	FRANCHISEES	LIENHOLDER	REORGANIZATIONAL	SUBLICENSEABLE
BANKRUPTCIES	FRANCHISER	LIENHOLDERS	REORGANIZATIONS	SUBLICENSED
BANKRUPTCY	FRANCHISERS	LIENS	REORGANIZE	SUBLICENSEE
BANKRUPTED	FRANCHISES	LIQUIDATE	REORGANIZED	SUBLICENSEES
CARRYBACK	FRANCHISING	LIQUIDATED	REORGANIZES	SUBLICENSES
CARRYBACKS	FRANCHISOR	LIQUIDATES	REORGANIZING	SUBLICENSING
CARRYFORWARD	FRANCHISORS	LIQUIDATING	REPATRIATE	SUBLICENSOR
CARRYFORWARDS	FUTURES	LIQUIDATION	REPATRIATED	SUBSIDIARIES
COLLABORATE	GLOBAL	LIQUIDATIONS	REPATRIATES	SUBSIDIARY
COLLABORATED	GLOBALIZATION	LIQUIDATOR	REPATRIATING	SUBSIDIES
COLLABORATES	GLOBALIZE	LIQUIDATORS	REPATRIATION	SUBSIDING
COLLABORATING	GLOBALIZED	LITIGATE	REPATRIATIONS	SUBSIDIZATION
COLLABORATION	GLOBALIZING	LITIGATED	RESTRUCTURE	SUBSIDIZE
COLLABORATIONS	GLOBALLY	LITIGATES	RESTRUCTURED	SUBSIDIZED
COLLABORATIVE	HEDGE	LITIGATING	RESTRUCTURES	SUBSIDIZERS
COLLABORATIVELY	HEDGED	LITIGATION	RESTRUCTURING	SUBSIDIZES
COLLABORATOR	HEDGES	LITIGATIONS	RESTRUCTURINGS	SUBSIDIZING
COLLABORATORS	HEDGING	LITIGIOUS	REVALUATION	SUBSIDY
COLLATERAL	IMBEDDED	MERGE	REVALUATIONS	SUBTENANCIES
COLLATERALIZATION	INFRINGE	MERGED	REVALUE	SUBTENANCY
COLLATERALIZE	INFRINGED	MERGER	REVALUED	SUBTENANT
COLLATERALIZED	INFRINGEMENT	MERGERS	REVALUES	SUBTENANTS
COLLATERALIZES	INFRINGEMENTS	MERGES	REVALUING	SWAP
COLLATERALIZING	INFRINGER	MERGING	REVOCABILITY	SWAPS
COLLATERALS	INFRINGERS	NATIONALIZATION	REVOCABLE	SWAPTION
COMPLEX	INFRINGES	NATIONALIZATIONS	REVOCATION	SWAPTIONS
COMPLEXITIES	INFRINGING	NATIONALIZE	REVOCATIONS	TAKEOVER
COMPLEXITY	INSOLVENCIES	NATIONALIZED	REVOKE	TAKEOVERS
COMPLEXLY	INSOLVENCY	NATIONALIZING	REVOKED	TRADEMARK
CONGLOMERATE	INSOLVENT	NONMARKETABLE	REVOKES	TRADEMARKED
CONGLOMERATES	INTANGIBLE	OUTSOURCE	REVOKING	TRADEMARKING
CONTINGENCIES	INTANGIBLES	OUTSOURCED	ROYALTIES	TRADEMARKS
CONTINGENCY	INTERCONNECT	OUTSOURCER	ROYALTY	UNEXERCISABLE
CONTINGENT	INTERCONNECTED	OUTSOURCERS	SECURITIZABLE	UNEXERCISED
CONTINGENTLY	INTERCONNECTEDNESS	OUTSOURCES	SECURITIZATION	UNRECOGNIZED
CONTRACT	INTERCONNECTING	OUTSOURCING	SECURITIZATIONS	UNREMITTED
CONTRACTED	INTERCONNECTION	PARTNER	SECURITIZE	UNREPATRIATED
CONTRACTHOLDER	INTERCONNECTIONS	PARTNERED	SECURITIZED	VENTURE
CONTRACTHOLDERS	INTERCONNECTS	PARTNERING	SECURITIZER	VENTURES
CONTRACTING	INTERNATIONAL	PARTNERS	SECURITIZERS	WARRANTEES
CONTRACTS	INTERNATIONALIZATION	PARTNERSHIP	SECURITIZES	WARRANTIED
CONTRACTUAL	INTERNATIONALLY	PARTNERSHIPS	SECURITIZING	WARRANTIES
CONTRACTUALLY	LAWSUIT	PATENT	SEGMENT	WARRANTING
CONTRACTUALS	LAWSUITS	PATENTABILITY	SEGMENTAL	WARRANTOR
CONTRACTURAL	LEASABLE	PATENTABLE	SEGMENTATION	WARRANTY
CONVERSION	LEASE	PATENTED	SEGMENTATIONS	WORLDWIDE
CONVERSIONS	LEASEABLE	PATENTEE	SEGMENTED	