A NATURAL LANGUAGE PROCESSING APPROACH TO IMPROVE DEMAND FORECASTING IN LONG SUPPLY CHAINS

by

William W.J. Teo

B.Acc., Singapore Management University, 2013

SUBMITTED TO THE PROGRAM IN SUPPLY CHAIN MANAGEMENT
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ENGINEERING IN SUPPLY CHAIN MANAGEMENT
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2020

Signature of Author:

_____

Department of Supply Chain Management
May 8, 2020

Certified by:

_____

Dr. Tugba Efendigil
Research Scientist
Thesis Advisor

Accepted by:

_____

Prof. Yossi Sheffi
Director, Center for Transportation and Logistics
Elisha Gray II Professor of Engineering Systems
Professor, Civil and Environmental Engineering

A Natural Language Processing Approach to Improve Demand Forecasting in Long Supply Chains

by

William W.J. Teo

Submitted to the Program in Supply Chain Management
on May 8, 2020 in Partial Fulfillment of the
Requirements for the Degree of Master of Engineering in Supply Chain Management

ABSTRACT

Information sharing is one of the established approaches to improve demand forecasting and reduce the bullwhip effect, but it is infeasible to do so effectively in a long supply chain. Using the polystyrene industry as a case study, this thesis explores the usage of modern natural language processing (NLP) techniques in a deep learning model, known as NEMO, to forecast the demand of a commodity — without requiring downstream companies to share information. In addition, this thesis compares the effectiveness of such an approach with other non-deep learning approaches, specifically an ARIMA model and a gradient boosting model, XGBoost, to demand forecasting. All three models returned large forecast errors. However, NEMO tracked the volatility of actual data better than the ARIMA model. NEMO also had better success in predicting demand than the XGBoost model, returning approximately 20% better Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) scores. This result suggests that NEMO can be improved with better data, but other issues, such as legality of text mining, need to be considered and addressed before NEMO can be used in day-to-day operations. In its current form, NEMO can be used alongside other forecasting models and provide invaluable information about upcoming demand volatility.

Thesis Advisor: Dr. Tugba Efendigil
Title: Research Scientist

Acknowledgments

In the course of writing this thesis, I have received a great deal of support and encouragement. I am very grateful to my thesis advisor, Dr. Tugba Efendigil, for her valuable guidance throughout. Her feedback helped me to refine my research focus. I would also like to thank the company for sharing the research data.

I am indebted to my wife, for being so understanding and for taking care of our family when I was occupied with my studies. Lastly, I am thankful for my children, who provided me with much needed encouragement. Without them, this thesis would not have been possible.

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1. Introduction

Demand forecasting is a challenging process but given the benefits of having an accurate forecast, companies are always looking to improve their forecasting process. One of the more promising methods that has emerged recently is sentiment analysis, which uses natural language processing (NLP) techniques to obtain useful information from text. Business-to-consumer (B2C) companies use this method to gather opinions and sentiments about their products from social media and e-commerce websites. However, business-to-business (B2B) companies have been slow to adopt sentiment analysis as B2B products are less likely to be discussed on social media and have reviews posted on e-commerce websites. However, other textual sources also contain valuable information about the end consumer. This thesis proposes a new NLP-based forecasting model, known as NEMO, that uses alternative textual sources of information with a modern NLP approach to B2B sales forecasting, and the practicalities and limitations of doing so.

1.2. Motivation

Demand forecasting is especially challenging for B2B companies selling commodities, and even more so if they are situated near the beginning of a long supply chain. Giunipero and Aly Eltantawy (2004) define a long supply chain as one where "there are three or more supplier tiers and/or when the products are globally bought, processed, and/or transported (p. 706)." The length of a supply chain directly impacts the lead time between companies; and consequently, the amount of safety stock needed and quantity of information required to forecast demand accurately (F. Chen, Drezner, Ryan, & Simchi-Levi, 2000). Inaccurate demand forecasts can cost companies millions of dollars in unsold inventories or unmet demand.

The field of NLP has seen rapid advancements since 2017, which allowed for more accurate downstream tasks such as sentiment analysis or text classification. This thesis investigates how can modern NLP techniques be used in the context of a B2B company selling commodities in a long supply chain. It is believed that such B2B companies, located far upstream in a supply chain, can improve their demand forecasting by employing modern NLP techniques to extract information from news articles.

Another area of interest for this thesis is comparing the performance of such an NLP-based forecasting technique to other forecasting techniques. The expectation is that by using these modern NLP techniques, B2B companies should be able to obtain up-to-date information about their end consumers and incorporate these information into their demand forecasts, making NLP-based forecasts more up-to-date and more accurate than forecasts derived from other techniques.

The above research topics will be formalized as research questions and hypotheses in Section 3.1.

1.3. Summary of Approach

This thesis conducts exploratory research using a petrochemical company as a case study. The petrochemical company produces and ships polystyrene pellets worldwide. As a B2B commodity, polystyrene has a wide range of application, from food packaging to household appliances. As seen in Figure 1, it also has a long supply chain, as polystyrene has to be processed, molded, repacked and distributed by various companies; before reaching the end consumer as plastic components in various consumer products, such as automobiles or televisions.

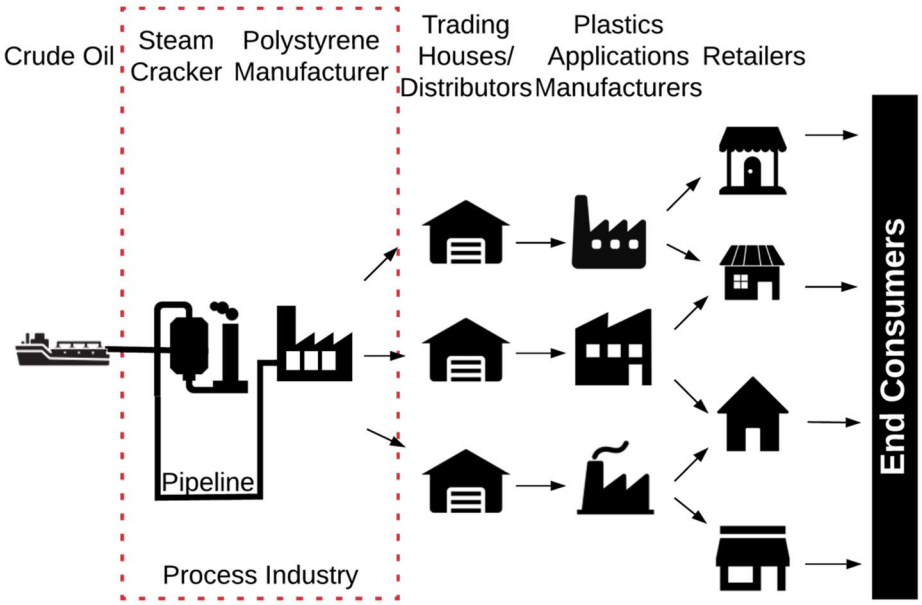


Figure 1. The Supply Chain of Polystyrene, a B2B Commodity

Five years of textual data are gathered because it takes a significant amount of manual effort to gather and process the textual data. Shynkevich, McGinnity, Coleman, and Belatreche (2015) found that using relevant news articles on targeted stocks, classified using the Global Industry Classification Standard (GICS), improved the prediction performance of their financial forecasting models. A similar approach is used in the selection of textual sources for NLP processing. At the same time, five years of empirical data, consisting of both polystyrene sales data and its related price indices, are obtained from the sponsor company. Sales quantity is used as a proxy for demand data because proper records of demand data were unavailable. For the purposes of this thesis, terms "sales" and "demand" will be used interchangeably except in situations where a clear distinction between the two is required.

The data are then processed and split into textual data and tabular data. Feature engineering is also carried out on the tabular data to generate additional variables that better represent the underlying data. The tabular data are then used as input into a neural

network, where categorical variables are converted into embeddings to capture the relationships between categories.

The textual data are then processed using modern NLP techniques. Since the introduction of transfer learning to NLP in 2017 (Vaswani et al., 2017), there has been a profusion of pre-trained language models, such as Google's Bidirectional Encoder Representations from Transformers (BERT) (Devlin, Chang, Lee, & Toutanova, 2018), which attained state-of-the-art results in a number of NLP tasks. NEMO uses one of these recently developed pre-trained language models to enhance the representation of textual data as numerical vectors.

After vectorization of the textual data, the vectors are combined with the output of the previous neural network and passed into another neural network, which predicts the demand quantity. The sales quantity and tabular data are also used to develop an autoregressive integrated moving average (ARIMA) model and a gradient boosting model respectively, and their results will be used to compare against NEMO's. All models are trained and evaluated using walk-forward cross-validation using the appropriate error measures.

The results from the above approach are then analyzed and evaluated against the research questions and hypotheses.

1.4. Gap in Existing Research

While NLP techniques have been used in forecasting, they mostly forecast stock prices or demand for end consumer goods. The most common NLP technique used to do so is sentiment analysis. However, existing research on using sentiment analysis to forecast demand in a supply chain context is extremely limited as sentiment analysis itself is a relatively new field. Wood, Reiners, and Srivastava (2013, 2014, 2015, 2016) have done

the most research in using sentiment analysis in such a context. In their latest paper, they developed a framework, shown in Figure 2, for the adoption of sentiment analysis in a firm, drawing on dual-process theory to conceptualize how information should be used in making decisions and how reflecting on the decision outcomes can be used to improve supply chain performance in a virtuous cycle. However, they also highlighted several limitations and challenges with such an approach (Wood et al., 2016).

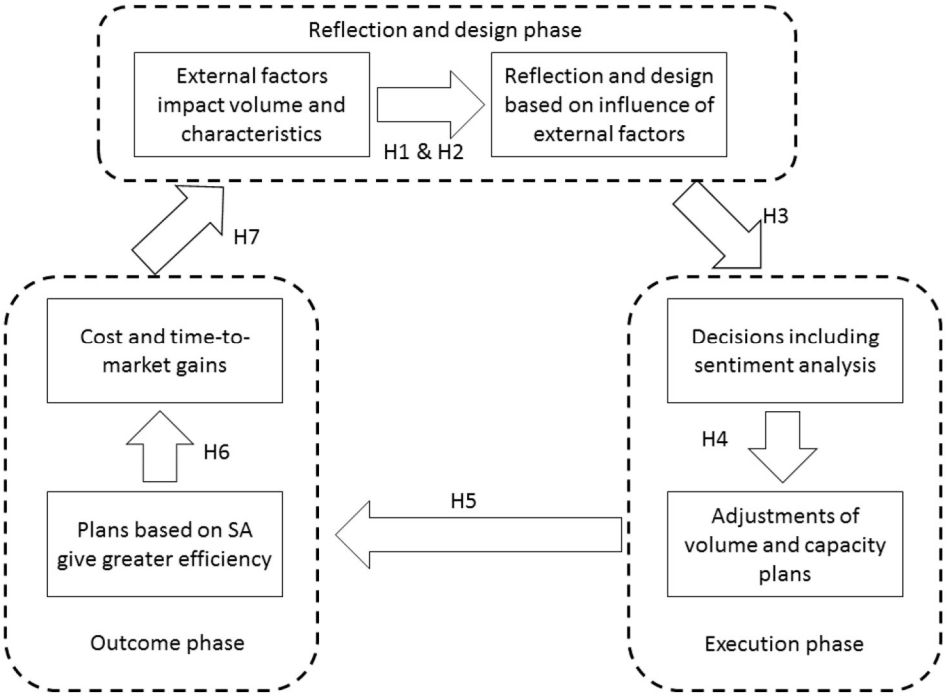


Figure 2. A framework for the adoption of sentiment analysis in a firm (Wood et al., 2016).

One noteworthy limitation they acknowledged is that sentiment analysis is "unlikely to be useful in B2B markets or with B2B products" as consumers are unlikely to express their opinions about such products on social media (Wood et al., 2016). They also stopped short of applying sentiment analysis to forecast demand in an actual case study. Furthermore, their papers predated the usage of transfer learning in NLP and hence did not take into account the impact of recently developed pre-trained language models and other modern

NLP techniques has on performance. While the approach proposed by this thesis does not use sentiment analysis directly, the NLP techniques used are similar and can be easily adapted to handle sentiment analysis too.

Another area that is unexplored by existing research is the derivation of useful information for forecasting from long formal text. Most of the existing research that used NLP techniques in forecasting focused only on relatively short length of text, such as Tweets or news headlines. These short pieces of text are relatively easy to interpret and incorporate into forecasting models. A long text document such as a news article may have multiple points of view and is therefore much harder to use in forecasting.

These gaps will be addressed in this thesis.

2. Literature Review

This chapter will review the existing literature on how NLP-based techniques can be used to forecast the demand of B2B companies selling commodities, especially chemical commodities; and the accuracy of different approaches in demand forecasting.

This literature review is broken down into three main sections: demand forecasting, NLP and industry. As this review will cover a broad spectrum of disciplines, only the most relevant and important areas in each field will be discussed. In the first section, the importance of demand forecasting and a conceptual framework to discuss supply chain forecasting will be introduced; followed by a survey of different methods used in forecasting and some challenges ahead for the field of forecasting. In the NLP section, a brief overview and recent developments in the field will be covered, before focusing on the usage of NLP techniques in forecasting and in supply chain management. Finally, relevant topics in the B2B, commodities and chemical commodities industries will be discussed in the last section to give context to the specific scope of this thesis.

2.1. Literature Review in Demand Forecasting

Demand forecasting is important to businesses and supply chains because it is the input to many important business decisions, such as the number of people to hire or the number of warehouses to build. Chase (2013) suggests that demand forecasting is important because operational processes takes time; and businesses "can no longer simply wait for demand to occur" and must instead "sense demand signals and shape future demand in anticipation of customer behavior (p. 31)."

Given the complexities of demand forecasting, it is useful to have a framework to organize the different aspects of a supply chain to consider in forecasting. One such framework is

proposed by Syntetos, Babai, Boylan, Kolassa, and Nikolopoulos (2016), as seen in Figure 3 and further elaborated in the sections below.

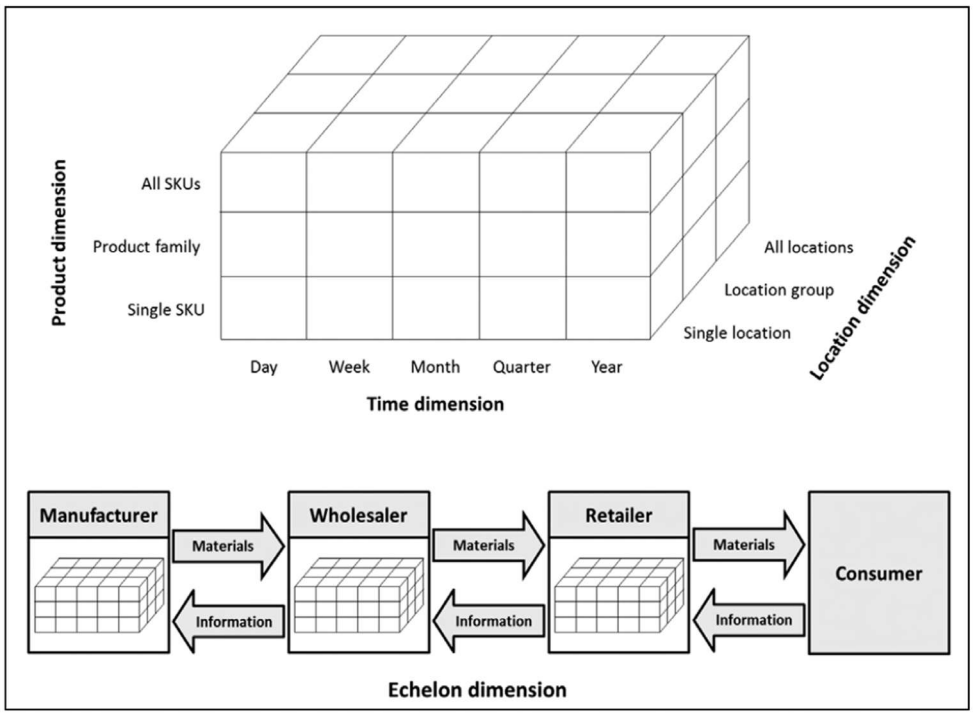


Figure 3. Theoretical framework of a supply chain structure informed by requirements of forecasting (Syntetos et al., 2016)

- Length

The length of a supply chain is captured by the echelon dimension in the framework proposed by Syntetos et al. and characterizes the flows of materials and information along the length of a supply chain. It becomes progressively harder to coordinate these flows the longer supply chains become (Syntetos et al., 2016), which is one of the main issues that this thesis addresses. Another issue is the "bullwhip effect," which is the amplification of demand variance along a supply chain as one moves further upstream.

- Depth

The depth of a supply chain is captured by both the location and product dimensions in Syntetos et al.'s framework. As forecasting "is a hierarchical process informing various

levels of decision making," different hierarchical elements, such as whether to aggregate SKUs or locations, are considered when capturing depth.

- Time

The time dimension is important for any forecasting problem and includes elements such as time buckets and forecast horizons (Syntetos et al., 2016).

2.1.1. Classifying Demand Forecasting Methods

Demand forecasting has many approaches, which can be classified in several ways. The approaches can broadly be divided into subjective and objective methods; and can be further broken down into judgmental and experimental approaches for the former and time series and causal approaches for the latter (Caplice & Sheffi, 2006a). Green and Armstrong (2012), in his review of demand forecasting methodologies, divided forecasting methods into judgmental and statistical approaches, as shown in Figure 4, the distinction being the availability of data.

2.1.1.1 Judgmental Methods

Judgmental methods rely on human judgment to forecast future trends. Judgmental forecasting is inevitable in many cases, such as a new product launch or no historical precedent (Hyndman & Athanasopoulos, 2018, p. 83). Green and Armstrong (2012) argues that the key to judgmental forecasting is to impose structure on judgments, using techniques like surveys and simulated interactions and to abstain from methods which have not been proven to be efficient, such as unaided judgments and focus groups.

2.1.1.2. Quantitative Methods

Quantitative forecasting uses data to forecast future trends, usually through statistical methods such as time series models or causal models, as described in the sections below.

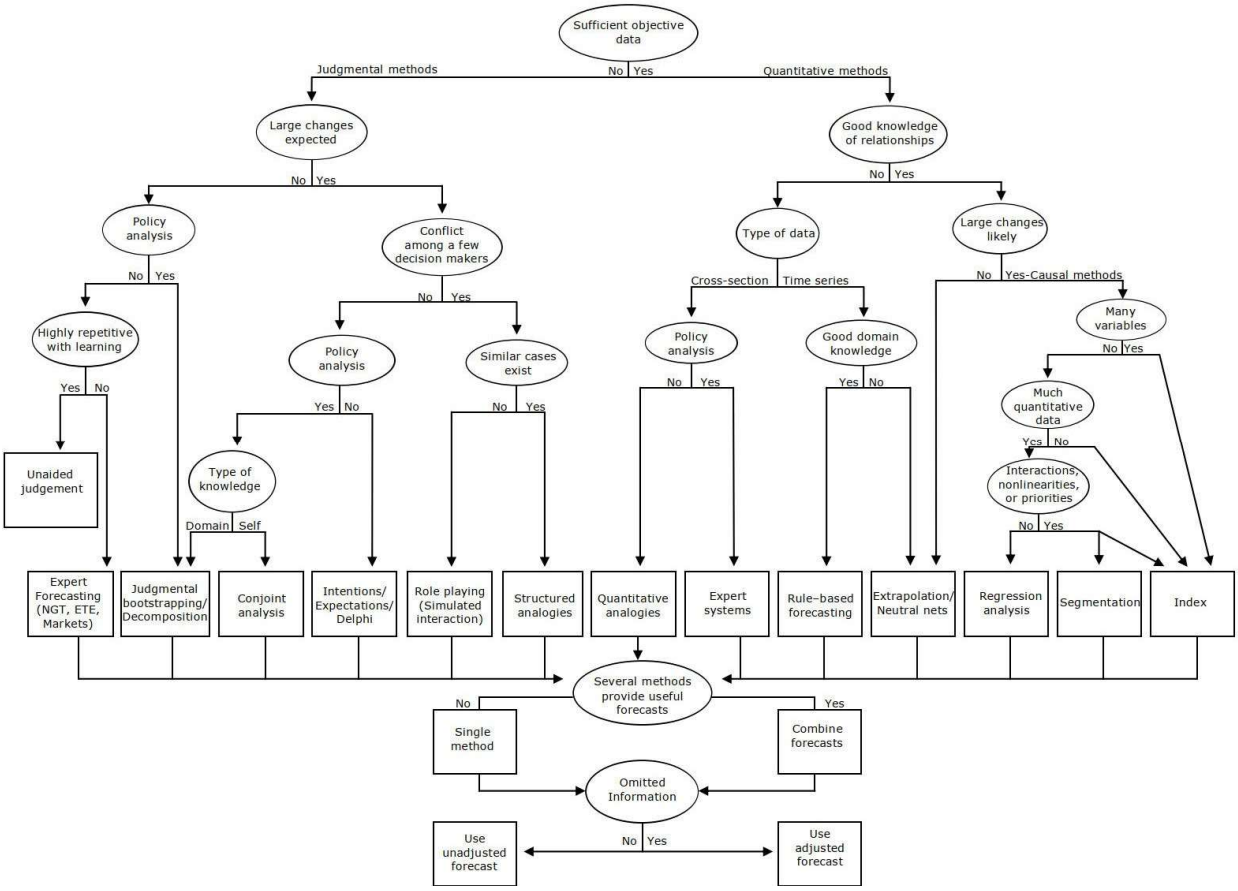


Figure 4. A flowchart classifying different forecasting methodologies. (Green & Armstrong, 2010)

2.1.1.2.1. Time Series Methods

Time series methods can be considered "traditional" methods as they have been around since the 1920s, when the statisticians Slutsky and Yule classified time series into autoregressive (AR), moving average (MA) or a combined autoregressive moving average (ARMA) processes (Nerlove & Diebold, 1990). In the 1970s, fellow statisticians Box and Jerkins further developed methods to estimate maximum likelihood and apply them to forecasting. Hence, ARIMA models are also known as Box-Jenkins models (Harvey, 1990). ARIMA models are heavily used in forecasting research. For example, issues underpinning the length and depth aspects of Syntetos et al.'s framework assume that "the underlying demand structure may be represented by an ARIMA form" (Syntetos et

al., 2016). Variations of the ARIMA model, such as MA or simple exponential smoothing (SES), form the bulk of forecasting methods used in practice today (Weller & Crone, 2012). For further reference, refer to the appendices in Syntetos et al. (2016)'s paper for a list of literature on the different types of statistical-based methods used in demand forecasting.

2.1.1.2.2. Causal Methods

Causal methods identify external explanatory variables that are highly correlated with demand and try to predict future demand using them (Caplice & Sheffi, 2006b). A common method to do so is through the usage of regression models; however existing knowledge and conceptual understanding rather than statistical fit are used in the selection of variables (Green & Armstrong, 2012).

Syntetos et al. (2016) did not cover causal methods in their paper but acknowledged that with the current popularity of Big Data, practitioners are trying to incorporate causal effects such as weather and social media in their forecasts. They pointed out that little research has been carried out to determine the extent to which explanatory variables forecast future trends in the field of supply chain management, although parallel developments in other disciplines such as marketing analytics exist; and that integration of ideas between disciplines will go towards bridging the gap between forecasting theory and practice.

2.1.1.2.3. Machine Learning Methods

Apart from traditional methods, newer artificial intelligence (AI) and machine learning forecasting methods, such as neural networks and support vector machines (SVMs), have gained popularity in recent years with the advent of increased processing power and Big Data. These techniques differ from statistical techniques as they can learn from the data

without any explicit rules. For example, neural networks are able to learn and model the non-linear patterns of intermittent demand (Mitrea, Lee, & Wu, 2009).

Several studies compared the performance of such techniques against traditional forecasting methods, but existing research indicated that is still premature to form a conclusion about which is a better approach. For example, Mitrea et al. (2009) compared the performance of neural networks against MA and ARIMA techniques in forecasting the demand for refrigeration compressors and concluded that neural networks performed better than traditional techniques. Carbonneau, Laframboise, and Vahidov (2008) compared the performance of several machine learning techniques, including neural networks and support vector machines, to traditional forecasting techniques, such as ARIMA variants and linear regression, in forecasting the demand of an extended supply chain. They found out that, although machine learning techniques had better accuracy, they did not offer a large improvement in accuracy over linear regression. They also pointed out that the marginal gain in accuracy of machine learning techniques must be weighed against the lower cost of adopting the simpler linear regression model.

In the more general field of forecasting, the popular Makridakis Competitions or M-Competitions are regularly held to evaluate the accuracy of different forecasting methods using a wide variety of time series data across various domains. An offshoot competition of the third edition of the M-Competitions, NN3, specifically assessed the forecasting accuracy of neural networks and other computational intelligence (CI) methods, such as K-nearest-neighbors or support vector regression, against other statistical methods. The results indicated that the non-statistical methods are comparable but still unable to outperform statistical methods (Crone, Hibon, & Nikolopoulos, 2011). The latest edition of the M-Competitions, M4, in 2018 was extended to all forecasting methods, including

machine learning methods, and involved the prediction of over 100,000 time series. Here, again the results indicated that submissions using purely machine learning methods underperformed other forecasting methods and 12 of the 17 most accurate methods comprised of combinations of mostly statistical methods (Makridakis, Spiliotis, & Assimakopoulos, 2018).

2.1.2. Combining Forecasting Methods

In practice, different types of forecasting methods are often combined. Reasons for combining forecasts include to increase overall forecasting accuracy and to eliminate bias (Chase, 2013). For example, the best forecasting method in the M4 competition was a hybrid of both statistical and machine learning methods (Makridakis et al., 2018).

Another reason is that different types of forecasts have their ideal use cases and a combination of forecasts is necessary in mixed cases. For example, in terms of the time dimension in Syntetos et al.'s framework, as seen in Figure 5, statistical forecasting methods are preferable for applications that have a short forecast horizon and long demand histories. The exact opposite is required for applications that require judgmental forecasting, such as a new product launch. The middle case is where judgmental forecasts are combined with statistical forecasting, which is also known as judgmental adjustments. Although commonly practiced in industry, very little academic research, mostly limited to empirical case studies, analyzed the effectiveness of such forecasts; as it is difficult to incorporate judgmental information into theoretical models (Syntetos et al., 2016). However, the few existing studies suggest that combining judgmental and quantitative forecasts are effective, especially if the judgmental adjustment were carried out in a systematic manner (Green & Armstrong, 2012).

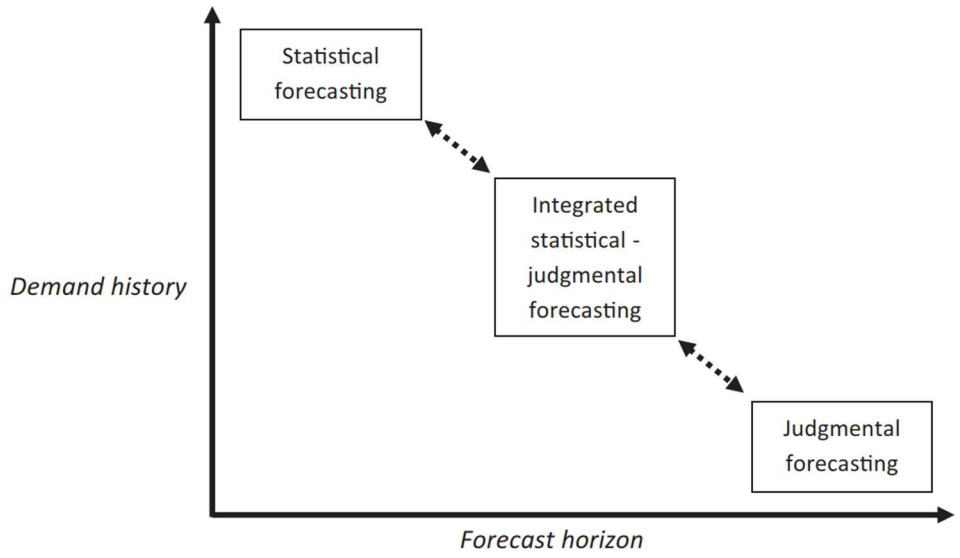


Figure 5. Link between judgmental forecasting and statistical forecasting (Syntetos et al., 2016)

2.1.3. Current Gaps and Challenges

In an editorial on the current state of supply chain forecasting, Boone, Boylan, Fildes, Ganeshan, and Sanders (2019) suggested to the forecasting community that, rather than trying to improve on existing time series models, they should expand their research agenda to explore the usage of Big Data and machine learning techniques; as "the role of artificial intelligence and machine learning methods in supply chain forecasting remains under-explored: the benefits and pitfalls of AI are not understood well in this context." Syntetos et al. (2016) also opined a similar view and highlighted "How to make effective use of information contained in social media?" is an important issue for further research.

2.2. Literature Review in Natural Language Processing

NLP is a collection of techniques that enables computers to interpret human language (Eisenstein, 2019, p. 1). It draws from many fields, most notably linguistics and computer science. A important subfield is sentiment analysis, which refers to the uncovering, extraction and classification of opinions, feelings and beliefs found in unstructured data

(Liu & Zhang, 2012). It can be useful in a variety of applications, from market intelligence to brand monitoring (Ravi & Ravi, 2015). It is still a relatively new field and only came into practical application recently with advancements in computing power and Big Data (Wood et al., 2014).

It is important to study the developments in the field of NLP because it touches upon issues and challenges in making sense of textual data, such as: "How to break down the text into a size suitable for processing?" Or "How to capture and represent the context found in sentences and paragraphs?" While an in-depth discussion of the history of NLP techniques is beyond the scope of this review, recent developments and their implications will be discussed in the following sections.

2.2.1. Traditional Approaches

Traditionally, parsing of text was done through a rule-based approach, first put forward by Chomsky (1957) in 1957. In the 1990s, statistical approaches became popular. It involved preprocessing and tokenizing textual data using statistical methods like term frequency-inverse document frequency (TF-IDF) or bag-of-words, before applying the tokenized vectors to the various NLP applications such as text classification or topic modeling (Bengfort, Bilbro, & Ojeda, 2018).

Some limitations of the above deterministic approaches to word vectors are that they do not differentiate between ordering of words and are unable to capture idiomatic context (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), which makes them not ideal for sentiment analysis purposes.

2.2.2. Deep Learning Approaches

Since the 2010s, there have been rapid advancements made in the field of NLP brought about by the resurgence of neural networks' popularity with researchers with the increase

in accessibility to computing power. Different variations of neural networks, such as long short-term memory (LSTM) networks, gated recurrent unit (GRUs), as well as new architectures of neural networks, such as the encoder-decoder architecture, attention mechanism or transformers, were introduced in rapid succession in the last few years (Lane, Hapke, & Howard, 2019). The most important and relevant developments are discussed in the sections below:

2.2.2.1. Word Embeddings

Instead of modeling words in a deterministic manner, word embeddings, such as word2vec developed by Google researchers (Mikolov et al., 2013) and GloVe developed by Stanford University researchers (Pennington, Socher, & Manning, 2014), are probabilistic approaches that can model context. This represented a huge leap forward in the field of NLP as subtle problems like semantic queries and word analogies are now possible (Lane et al., 2019). The classic example of a word analogy is "King - Man + Woman = Queen," where the underlying gender relationship is captured by word embeddings (MIT Technological Review, 2015). This had huge implications for NLP as it is now possible to capture and compare the semantic information encoded in different passages of text.

2.2.2.2. Transfer Learning

Alongside the development of word embeddings, another important development is that of transfer learning. Transfer learning refers to the situation where previous learning done in a context can be used to improve generalization in another context (Goodfellow, Bengio, & Courville, 2016, p. 526). It has been used in many other fields, most notably in computer vision; where Oquab, Bottou, Laptev, and Sivic (2014) managed to apply image representations trained on the ImageNet dataset to other visual recognition applications with minimal retraining.

Many different types of transfer learning in NLP exist, among which sequential transfer learning has contributed to the biggest gain in accuracy of sentiment classification to date. The general idea is to pretrain on a large corpus of text, such as Wikipedia, before adapting the pre-trained model to another supervised task (Ruder, 2019). There have been a number of pre-trained language models, each having different architectures, released in the last few years; such as Bidirectional Encoder Representations from Transformers (BERT) by Google (Devlin et al., 2018), Universal Language Model Fine-Tuning (ULMFiT) by fastai (Howard & Ruder, 2018) and XLNet by Google (Yang et al., 2019).

Table 1. A comparison of different pre-trained models' accuracy in sentiment analysis on the IMDb dataset

Model	Accuracy	Paper
XLNet (Yang et al., 2019)	96.21%	XLNet: Generalized Autoregressive Pretraining for Language Understanding
BERT-large with In-Task Pre-Training (Sun, Qiu, Xu, & Huang, 2019)	95.79%	How to Fine-Tune BERT for Text Classification?
BERT-base with In-Task Pre- Training (Sun et al., 2019)	95.63%	How to Fine-Tune BERT for Text Classification?
ULMFiT (Howard & Ruder, 2018)	95.4%	Universal Language Model Fine-tuning for Text Classification

As seen from Table 1, pre-trained language models are very accurate in sentiment analysis and have been increasing in accuracy.

2.2.3. Usage of NLP techniques in Forecasting

While rapid advancements have been made in the field of NLP, most of the existing research on the usage of NLP techniques for forecasting is limited to traditional approaches and has not incorporated probabilistic word embeddings or transfer learning

in their methodologies. They are also mostly limited to either price forecasts or consumer products.

Table 2. A survey of existing literature on NLP techniques used in forecasting

Paper	NLP Techniques	Forecast Target
Product Sales Forecasting Using Online Reviews and Historical Sales Data: A Method Combining the Bass Model and Sentiment Analysis (Fan, Che, & Chen, 2017)	Word frequency, Naive Bayes classifier	Car sales
Stock Trend Prediction Using News Sentiment Analysis (Kalyani, Bharathi, & Jyothi, 2016)	Bag-of-words, polarity dictionary	Stock prices
Increasing the Explanatory Power of Investor Sentiment Analysis for Commodities in Online Media (Klein, Riekert, Kirilov, & Leukel, 2018)	Unigrams, SVM classifier	Commodity futures prices
Predicting Vehicle Sales by Sentiment Analysis of Twitter Data and Stock Market Values (Pai & Liu, 2018)	Polarity scoring	Car sales
Parallel Aspect-Oriented Sentiment Analysis for Sales Forecasting with Big Data (Lau, Zhang, & Xu, 2018)	Polarity scoring	Consumer products
Stock Price Prediction Based on Stock-Specific and Sub-Industry-Specific News Articles (Shynkevich et al., 2015)	Bag-of-words, polarity scoring	Stock prices
Text-Based Crude Oil Price Forecasting: A Deep Learning Approach (X. Li, Shang, & Wang, 2018)	Bag-of-words, polarity scoring	Crude oil prices

As seen from Table 2, existing research has not adopted the latest developments in the field of NLP. Given the high accuracy of using word embeddings and transfer learning in NLP, it would be worthwhile to explore the usage of such methods applied to forecasting. Another interesting observation is that the prediction of stock or commodity prices is usually done through textual analysis of news articles while the demand of consumer products is predicted using sentiment analysis of social media. To date, no study has been carried out on demand forecasting of commodities using NLP techniques, or whether this is achievable through the textual analysis of social media or news articles.

2.2.4. Usage of NLP techniques in Supply Chain Management

As mentioned in an earlier section, existing research on using NLP techniques in the field of supply chain management is extremely limited. The possibility of doing this kind of analysis, specifically through sentiment analysis, was proposed by both Wood et al. (2013) and Swain and Cao (2013) concurrently. Preliminary investigations by Swain and Cao (2013) showed that higher levels of social media activity by supply chain partners led to better supply chain performance.

Wood et al. (2015) suggested that through the usage of sentiment analysis of social media, it is possible for upstream firms to sense changes in market demand without relying on information shared by downstream firms. In their latest paper on the topic, Wood et al. (2016) formulated a framework for the usage of sentiment analysis in a firm, drawing on dual-process theory to conceptualize how information gained from sentiment analysis should be used in make supply chain decisions and how reflecting on the decision outcomes can be used to improve supply chain performance in a virtuous cycle. As an extension of their earlier paper, they also proposed that: by having early access to market data through the usage of sentiment analysis, it will help firms to respond quicker to market changes and counteract the bullwhip effect, especially for firms that are distant from the end market.

Lastly, Swain and Cao (2017) carried out an exploratory study and found support for a correlation between supply chain performance and amount of social media shared, where supply chain performance is measured by inventory turnover.

The papers cited above considered the sentiment analysis only of social media and were more focused on theorizing how to apply sentiment analysis to improve supply chain performance, not demand forecasting. Furthermore, Wood et al. (2016) even pointed out

that sentiment analysis is “unlikely to be useful in B2B markets or with B2B products” as consumers are unlikely to express their opinions about such products on social media.

2.3. Industrial Context

The following sections give a brief background to the industrial context of this thesis.

2.3.1. Business-to-Business

B2B refers to the sales conducted between businesses. A key issue of applying sentiment analysis to B2B sales is that B2B businesses typically perceive their products as “rational” and hence do not engage in as much branding (Beverland, Lindgreen, Napoli, Kotler, & Pfoertsch, 2007) and social media marketing as B2C businesses (Iankova, Davies, Archer-Brown, Marder, & Yau, 2019). This makes it a challenge to find opinions on B2B products online.

2.3.2. Commodities

Commodities are goods that standardized, undifferentiated, have a uniform price and often used in the production of other goods ("What Makes Something a Commodity?," 2017). A key difference between consumer products and commodities is that since commodities are so vital to commerce and have rather standardized prices; they are highly liquid and often traded on exchanges as a financial security, along with all the derivatives like futures and options ("What Makes Something a Commodity?," 2017). Cheng and Xiong (2014) call this the “financialization” of commodity markets and argue that this is a positive development as it promotes information discovery and that future prices are important demand signals of commodities. In addition, being a financial security, ample information can be found online in the form of news articles and price benchmarking agency reports (Johnson, 2018), which can be mined for sentiment analysis.

Chemical commodities are mass produced products and their sales are primarily driven by their selling price. Examples include plastic polymers such as polyethylene terephthalate (PET) or basic chemicals such as chlorine. Prices are volatile (Kannegiesser, Günther, van Beek, Grunow, & Habla, 2009) and often track a commodity benchmark such as that of ICIS or S&P Global Platts (Johnson, 2018).

2.3.3. Demand Forecasting in the Industry

Demand forecasting can be and have been applied in the B2B sales (Lackman, 2007) as well as commodities (Xu, Qi, & Hua, 2010). Forecasting the demand for chemical commodities is challenging because demand for chemical commodities is not autoregressive (Kannegiesser et al., 2009). Furthermore, chemical commodities sales are made up of spot and contract sales; and their demand patterns are different — spot demand does not need to be met while contract demand is fixed. (Kannegiesser et al., 2009). For these two reasons, Kannegiesser et al. (2009, p. 66) claim that "the classical approach towards demand forecasting does not apply to the considered chemical commodity business."

2.4. Summary

Given the wide range of topics covered by this thesis, this literature review has focused on only the most relevant and important areas in each relevant field. Demand forecasting is complicated because of the many different dimensions to consider, as seen in Syntetos et al.'s framework. Furthermore, a large variety of forecasting approaches exist, from judgmental methods to machine learning methods. Sentiment analysis is another new approach to forecasting, made possible with the rise of Big Data and recent advancements in NLP techniques, such as word embeddings and transfer learning. While there has been some research on using NLP techniques in forecasting, they were limited to either price

forecasts or demand of consumer products. In addition, very little existing literature on the usage of NLP techniques in supply chain management exists. Although it is hard to find sentiments of B2B commodities online, the "financialization" of commodity markets gives rise to other potential sources of information, such as analyst reports and news articles, for NLP-based forecasting. Lastly, research shows that the standard forecasting approach may not be suitable for forecasting the demand of chemical commodities.

This thesis will fill several knowledge gaps across different disciplines. An NLP-based forecasting approach, such as sentiment analysis, is not well-studied and defined. This thesis will also help to answer Boone et al. (2019)'s call to forecasting researchers to better understand the benefits and pitfalls of AI in supply chain forecasting and contribute to the scarce literature on the usage of NLP techniques in supply chain management. In addition, this thesis will use the latest NLP techniques, such as word embeddings and transfer learning, which displayed very high accuracy in sentiment analysis benchmarks. No existing paper has used such techniques to forecast demand before. Lastly, the author believes that this thesis is the first paper that attempts to use modern NLP techniques to forecast the demand of B2B companies selling commodities specifically. Hence, this thesis helps to establish the feasibility of such an approach and a method to do so.

3. Methodology

The primary objective of this thesis is to establish a method to extract the latent semantic information embedded in textual documents and use it to forecast the demand for polystyrene. Traditional NLP techniques, such as sentiment analysis approaches to forecasting, tend to calculate a sentiment score and use it as a variable to forecast demand (X. Li et al., 2018; Pai & Liu, 2018). In such approaches, a sentiment lexicon, which is a list of positive and negative sentiment words and phrases, is required to compute the sentiment score. However, human judgment is needed to determine what constitutes a 'positive' or 'negative' sentiment and decide the rules to calculate 'intensity' of the sentiment (Liu, 2015). The proposed approach bypasses the need for such sentiment scoring by using deep neural networks to learn the hidden relationships between the textual data and demand.

As seen from Figure 6, there are five phases in the proposed methodology: data retrieval, data cleaning, deep learning-based NLP, demand forecasting and evaluation. In the first two phases, data are collected and cleaned for modeling. In next phase, textual data are preprocessed and then converted into numerical representations or embeddings for forecasting. In the demand forecasting phase, the embeddings are then combined with other data before using neural networks to forecast demand. This NLP-based approach to demand forecasting, represented by the blue box in Figure 6, is NEMO. NEMO takes reference from Tremblay's model, which merges image, tabular and textual data to predict pets' adoption rates (Tremblay, 2019).

Two other forecasting methods, an ARIMA model and a gradient boosting model, are also used to forecast demand. The ARIMA model only uses the sales data for prediction, while

the gradient boosting model uses all of the tabular data for prediction. Finally, performances of the three models are evaluated in the last phase.

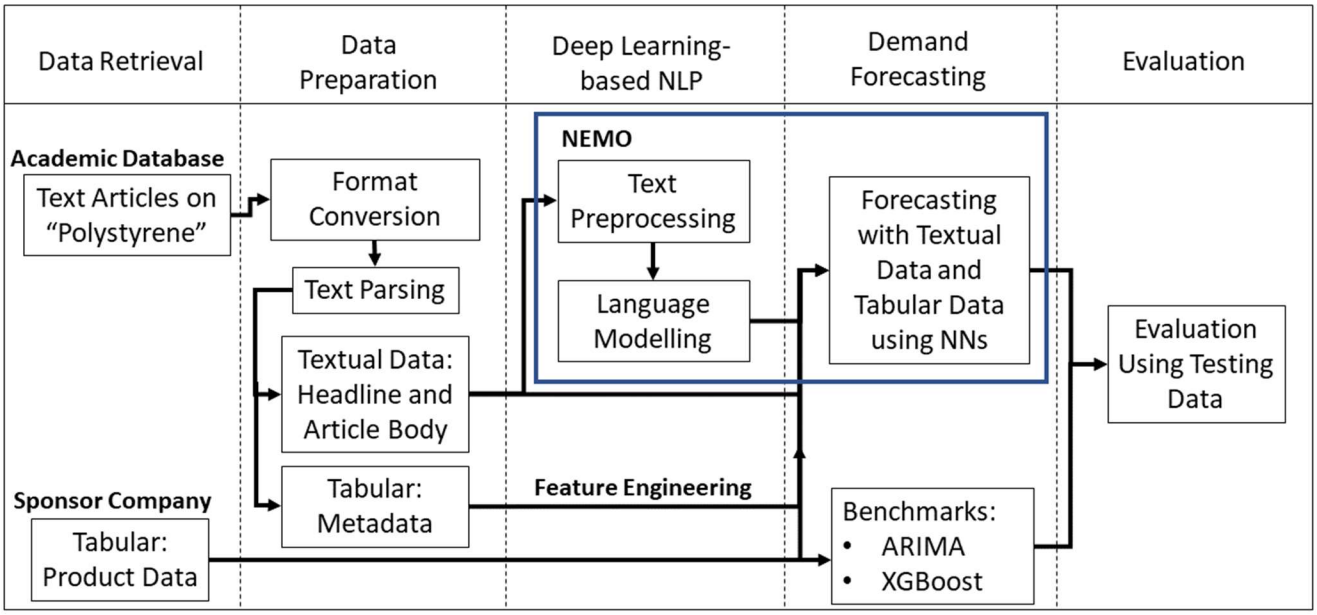


Figure 6. Proposed Methodology

3.1. Research Questions and Hypotheses

This thesis explores the possibility of bypassing the need for information gathering from downstream companies through the application of NLP techniques to information readily available online.

Research Question One: How can NLP techniques be applied to forecast the demand of B2B companies selling commodities in long supply chains?

For example, as discussed in previous chapters, sentiment analysis is typically used by B2C companies in purposes like market analysis or brand monitoring; as it is easier to capture consumer sentiments of a branded product on social media. In contrast, B2B products are usually not discussed on social media; and hence, it is difficult to apply sentiment analysis to them. However, sentiment analysis has been used on news articles

to forecast commodities prices such as gold (Smales, 2014) and crude oil (J. Li, Xu, Yu, & Tang, 2016).

Hypothesis One: News articles contain sufficient information to apply NLP techniques to forecast the demand of B2B commodities accurately.

Although this thesis is based on B2B commodities, the same NLP-based approach can be extended to forecast the demand of B2C products easily.

In addition, given the speed at which exogenous factors impact demand today, traditional forecasting methods are not able to incorporate such information in their forecasts as quickly as one would like.

Research Question Two: Does a deep learning NLP-based forecasting approach have greater forecast accuracy than non-deep learning forecasting approaches?

This thesis will also compare the performance of a deep learning NLP-based forecasting approach against other forecasting methods, such as an ARIMA model.

Hypothesis Two: NLP techniques can incorporate the latest exogenous factors in its forecasting model, making forecasts more up-to-date and more accurate.

3.2. Methodology Steps

3.2.1. Data Retrieval

Two types of data, textual data and empirical data, are collected in this phase. Textual data are obtained from an academic database called Nexis Uni, while empirical data are obtained from the sponsor company. Empirical data consist of both the sales data of polystyrene and its related price indices. A closer analysis of data is detailed in Section 4.1. For both textual and empirical data, five years of data, from the start of 2014 to the end of 2018, is gathered. However, the frequencies of the data vary: textual data are daily, price indices are weekly, and sales data are in monthly buckets. The relevant price indices

data and sales data are upsampled to match the daily frequency of the textual data using the Python library pandas.

3.2.2. Data Cleaning

In this phase, text documents downloaded from the database are cleaned up for NLP. The downloaded articles contain metadata such as industry and subject classifications. The articles are converted into an appropriate file format and parsed using Python. The headline and article contents are then extracted and stored separately from the metadata as textual data. Duplicates articles are not cleaned up because the number of articles may be a predictor of demand.

The metadata are then combined with the empirical data, undergoes feature engineering and one-hot encoding to create a tabular dataset. Tabular data are data that are found in tables and are made up of either categorical data or continuous data. All the categorical variables are transformed into embeddings, which allows multi-dimensional relationships between categories to be captured. Furthermore, time series data are also split into multiple categorical variables to capture such multi-dimensional relationships too. For example, time series data can be split into whether it is the first quarter of the year or the specific day of the week. One-hot encoding is simply classifying if the data are about a certain industry or company, encoding 1 if true and 0 if false.

3.2.3. Deep Learning-based Natural Language Processing

This thesis uses fastai, which builds upon the open source machine learning library PyTorch. fastai helps simplify the training of fast and accurate neural networks using modern best practices (Howard & Gugger, 2018), such as using one-cycle learning (Smith, 2018) to fit the neural network model.

3.2.3.1. Text Preprocessing

Minimal preprocessing is carried out on the textual data because traditional text preprocessing techniques, such as stopword removal or lemmatization, are unnecessary for deep learning and will in fact result in the loss of important semantic information.

fastai contains preprocessing functions for tokenization and numericalization, which are necessary to turn textual data into tokens and tokens into unique IDs respectively. These unique IDs are subsequently used to create the vocabulary of the language model. As every word requires a distinct row in the resulting neural network's weight matrix, the vocabulary is limited to 60,000 words to avoid having an overly large matrix to process later.

3.2.3.2. Language Modeling

This thesis uses transfer learning to reduce the time needed to train a language model that has deep contextualized representations of the underlying text. fastai uses a subset of English Wikipedia for its pretrained language model, which can be fine-tuned with textual data from the target domain. In the context of this thesis, the language model is fine-tuned with the articles downloaded from Nexis Uni containing the keyword "polystyrene." Since the language model is not used to predict demand directly, it is trained over the entire textual dataset using a recurrent neural network (RNN), which is the typical variant of neural networks used in NLP.

At this point, the language model will be able to predict the next word of any input phrase, using its knowledge of sentences and words represented by the embeddings in it. However, since the aim of this thesis is understanding textual data and its hidden sentiments, only the part of language model responsible for encoding the textual data into numerical vectors is saved for later usage. This part is also known as the "encoder."

3.2.4. Demand Forecasting

In this phase, the preprocessed data are used in three different models to predict demand.

3.2.4.1. Forecasting with Textual Data and Tabular Data

As seen from NEMO's architecture in Figure 7, tabular data and textual data are passed through several fully connected linear layers and a RNN respectively before concatenating them in a single tensor, passed through several more linear layers and finally trained from end-to-end to forecast demand. It is important to ensure that the textual data are processed with the same encoder from the language model trained previously in Section 3.2.3.2. The specific type of RNN used to process textual data is fastai's ASGD Weight-Dropped LSTM (AWD-LSTM), which incorporates several very effective regularization and optimization strategies in its implementation.

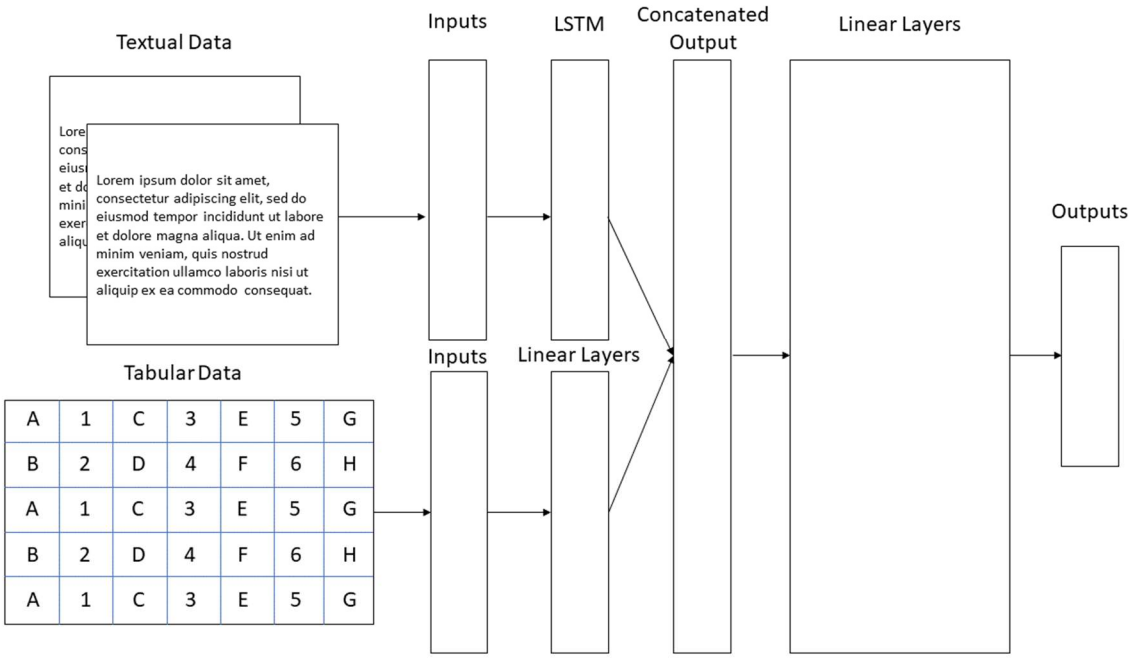


Figure 7. NEMO's Neural Network Architecture

The concatenated output is then passed through several more fully connected layers and trained end-to-end on the target output, which is the sales quantity for that month. As

this is essentially a regression problem, regression metrics are used as the loss functions to train the model. The regression metrics used are root mean square error (RMSE) and mean absolute error (MAE), which are calculated as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)}{n}}$$

(Eq. 1)

and

$$MAE = \frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|}{n}$$

(Eq. 2)

where $\hat{y}_t$ is the predicted value at time t and y_t is the actual value at time t, respectively. Having two versions of a model is useful because the two measures have different mathematical properties and hence purposes. The mathematical definitions of RMSE and MAE result in a model forecasting the mean when minimizing for RMSE, and a model forecasting the median when minimizing for MAE. When carrying out hierarchical or group forecasting, it is possible to aggregate forecasts by adding up forecasts of different groups. However, only forecasts of means can be added up (Hyndman & Athanasopoulos, 2018). On the other hand, a model that minimizes for MAE is robust to errors and is less likely to result in outliers (Chai & Draxler, 2014).

3.2.4.2. Statistical Forecasting

An ARIMA model, which is one of the two most widely used approaches in traditional statistical forecasting, is used to forecast demand in this section. As an ARIMA model has many different components, the following approach, which observes the

recommendations given by Hyndman and Athanasopoulos (2018, pp. 243-252), is used to select the order of the ARIMA model.

First, plot the training data to identify any unusual observations. Next, check whether the variance of the data is stable and apply a Box-Cox transformation if not. Then, apply unit-root tests to determine if any differencing is necessary to correct for seasonality and to make non-stationary data stationary. Hereafter, try out different ARIMA models to find a suitable ARIMA model. The criteria used to choose models are the lowest root mean square error (RMSE) and mean absolute error (MAE) scores using the validation set respectively. The residuals of the chosen models are then tested and checked by plotting the autocorrelation (ACF) plots of the residual and doing a portmanteau test of the residuals. If the residuals look like white noise, the chosen ARIMA model is used to forecast the demand in the testing set. The ARIMA models chosen are described in Section 4.2.2 of the Data and Results Chapter.

3.2.4.3. Machine Learning-based Forecasting

Apart from a statistical forecasting method, a machine learning-based forecasting method is also selected and trained on the tabular data for like-for-like comparison. Extreme gradient boosting (XGBoost) is chosen because of its good performance in practice, its suitability with categorical data and relatively fast computational times as compared to other machine learning techniques such as support vector regression. It is an ensemble of weak learners or trees and is generalized by optimizing of a differentiable objective function (T. Chen & Guestrin, 2016).

XGBoost has many parameters to adjust, including parameters like regularization parameters and depth of its trees. Initially, grid search and random search were employed to find the best parameters but the number of parameters to explore meant that model

tuning took a very long time. Subsequently, Bayesian optimization is used to select the parameters. Bayesian optimization is a search technique based on Bayes theorem, which makes the parameter optimization process very efficient (Snoek, Larochelle, & Adams, 2012).

Like NEMO and the ARIMA model, XGBoost uses RMSE and MAE as the error functions to train the models in cross-validation. However, as XGBoost requires a differentiable objective function, the pseudo-Huber loss is used as an approximation for the MAE instead when training the model.

3.2.5. Evaluation

As mentioned before, the models' performances are evaluated using walk-forward cross-validation with RMSE and MAE as the error measures. RMSE is widely used to compare performances between models because it gives a higher penalty to errors, making it an ideal measure to select the best model. MAE, as the average absolute difference between the predicted and actual value, is easier to interpret compared to RMSE. For the XGBoost model and NEMO, since the underlying data frequency is daily, their predictions are averaged to obtain a monthly prediction for evaluation.

Initially, the models were evaluated on a simple train-validate-test split. However, cross-validation — a more sophisticated evaluation method — is used subsequently as it "provides an almost unbiased estimate of the true error (Varma & Simon, 2006)." The models, ARIMA, XGBoost and NEMO, are carefully chosen to evaluate different forecasting approaches. The ARIMA model is a statistical model that only uses past observations of time series data in its forecasts. The XGBoost model is a highly regarded machine learning model and incorporates external information in its forecasts. NEMO is a deep learning model that uses NLP to extract information from long textual documents

and uses them, along with the same external variables used by the XGBoost model, to forecast demand. All three models are evaluated using time series cross-validation, specifically walk-forward cross-validation, where the training or validation set only contains data that occurs before the validation or test set respectively. The walk-forward cross-validation is used to avoid look ahead bias, which occurs when a model is trained on future information not yet available in the training period (Hyndman & Athanasopoulos, 2018). As seen in Figure 8, a year's worth of data is used for each dataset to account for any yearly seasonality effect, which results in three cross-validation folds. For each cross-validation fold, the models' hyperparameters are tuned using the validation set and the tuned models' performance is evaluated against the testing set. The models' final performance is the average of the error measure across three different folds.

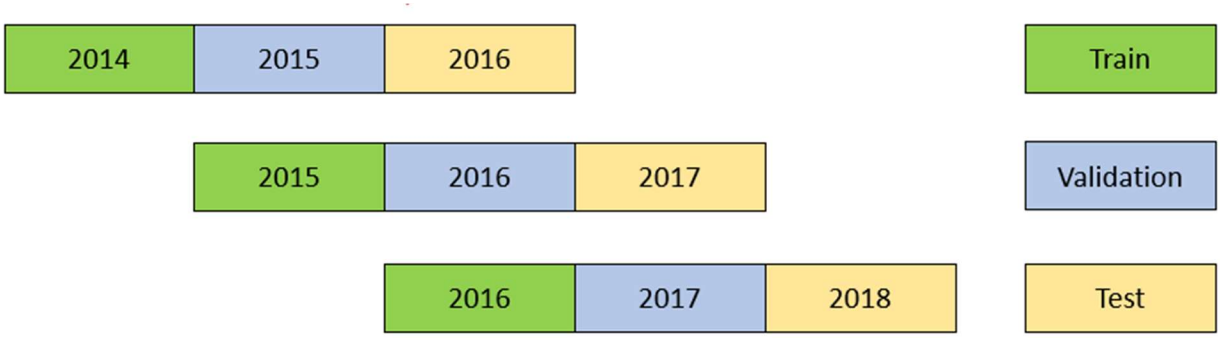


Figure 8. Walk-forward Cross-Validation

3.3. Summary

In this chapter, a five-step methodology is proposed, which outlines how a modern NLP-based deep learning model can be used to extract information from long textual documents to forecast demand and how will such a model be evaluated against other types of forecasting models. This methodology will help explore the research questions and the related hypotheses on using news articles to forecast the demand of B2B commodities and establish the accuracy and timeliness of NLP-based forecasting techniques.

4. Data and Results

The previous chapter laid out the methodology used to execute and evaluate three different forecasting models. In this chapter, both the input data and model results are described and evaluated. The first section will explore the data used in greater detail while the second section will look at the results obtained from the models and how to interpret them.

4.1. Data

Textual data are collected from Nexis Uni, an academic research database with more than 15,000 news, legal and business sources. Empirical data consisting of sales data for polystyrene and its related price indices are obtained from the sponsor company.

4.1.1. Textual Data

"Polystyrene" is used as the search keyword in Nexis Uni and results are further narrowed down by industry and language, using "Chemical" and "English" respectively. This is similar to Shynkevich, McGinnity, Coleman and Belatreche's (2015) successful approach of using industry-specific news articles in their stock price predictions.

4.1.1.1. Data Distribution

Five years of textual data are manually downloaded from Nexis Uni. As seen in Figure 9 and Table 3, distribution of the data is not uniform. A dip in textual data available can be observed from April 2015 to August 2015 and more data are available in 2017 and 2018. This anomaly is attributed to availability of articles in the database. The uneven distribution of data may affect the models' accuracies. This will be discussed further in Section 4.3.

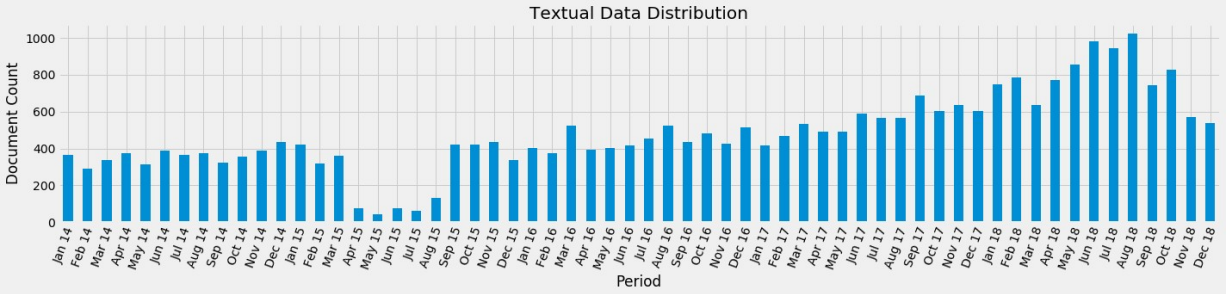


Figure 9. Time Series Plot of Textual Data Distribution

Table 3. Number of Articles per Year

Year	Number of Articles
2014	4,308
2015	3,102
2016	5,339
2017	6,657
2018	9,423

4.1.1.2. Publication Types and Subject Classification

The publication types differ greatly, varying from pricing reports to news reports about regulations banning the usage of plastics. However, as seen in Figure 10, the vast majority are news-related articles.

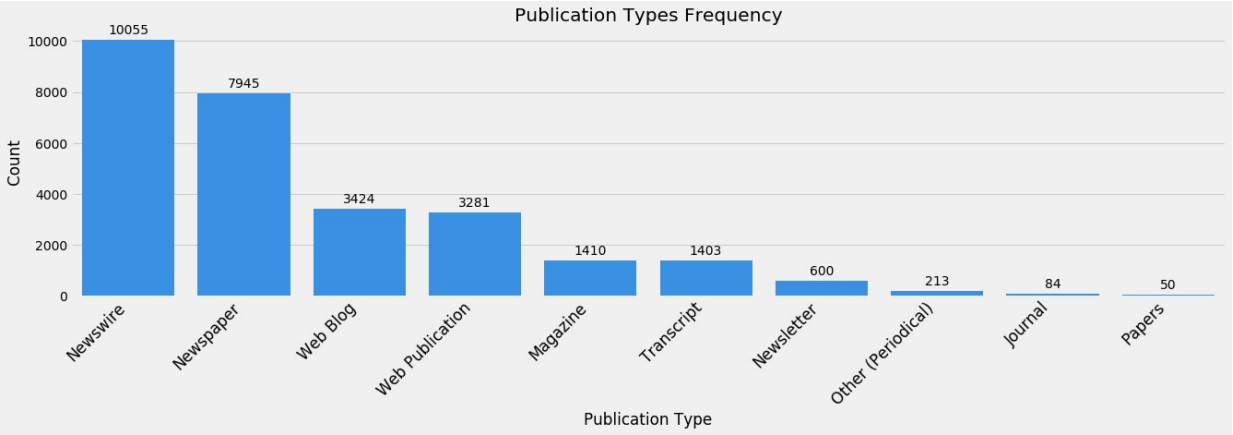


Figure 10. Frequency Plot of the Top 10 Publication Types

Nexis Uni classifies documents according to subjects using its own classification system. A document can have more than one subject classification. The top 20 subjects based on count frequency are shown in Figure 11. As expected, most of the data are documents on the plastic polymer industry and its trends.

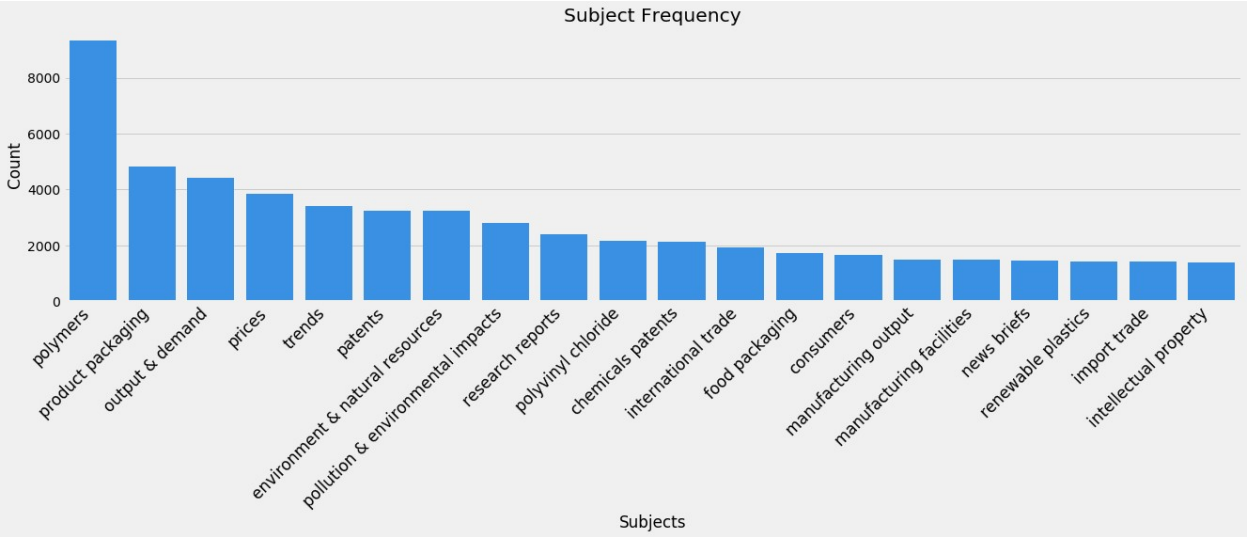


Figure 11. Frequency Plot of the Top 20 Subjects

4.1.1.3. Document Length

As seen from Table 4, the average document length is 775 words with a standard deviation of 3,450 words. As seen in Figure 12, the distribution is skewed by a few outliers with document lengths of over ten thousand words. However, the median document length is still fairly long at 471 words, and this is consistent with the type of textual documents collected, which are mainly news articles and pricing reports.

Table 4. Descriptive Statistics of Document Length

Mean	775
Standard Deviation	3,450
Minimum	18
25%	261
50%	471
75%	781
Maximum	188,180

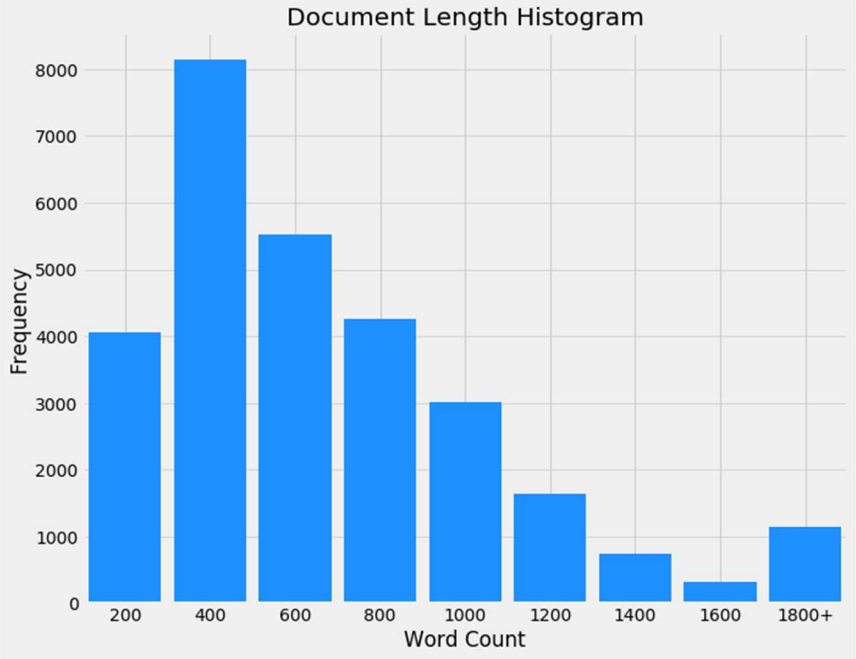


Figure 12. Histogram Plot of Document Length

4.1.2. Empirical Data

The two types of empirical data are described in the subsections below.

4.1.2.1. Price Indices

Price indices are obtained from major pricing reporting agencies, such as S&P Global Platts, and consist of benchmark prices for polystyrene and its related chemicals, such as benzene and styrene. Benchmark prices are classified according to regions with the relevant Incoterms, such as Free on Board (FOB) Korea or Cost and Freight (CFR) Hong Kong. Pricing reporting agencies obtain their data by inquiring prices from contacts in major market participants; therefore, for each price index, there is a minimum and maximum price (Johnson, 2018). A total of 32 price indices of various chemicals and from different regions of the world are taken into consideration.

While specific price indices used cannot be shared due to licensing agreements, the general price trends can be seen in Figure 13.

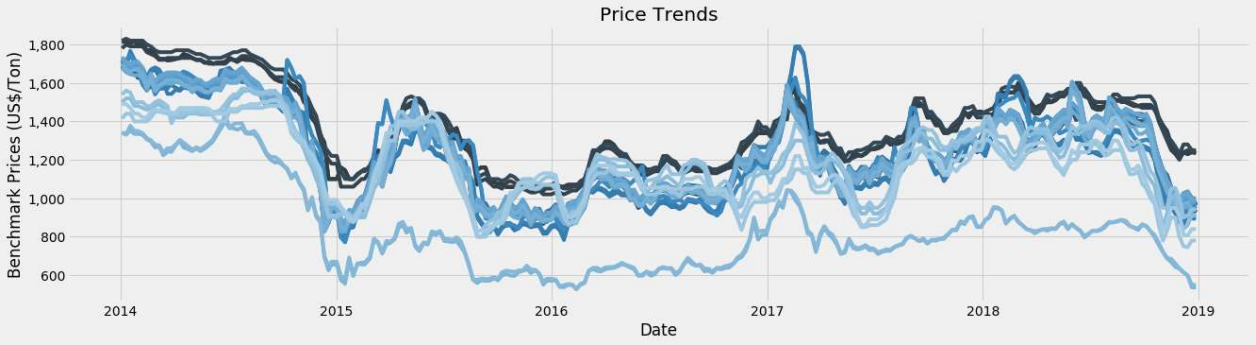


Figure 13. Time Series Plot of Various Regional Chemical Price Indices

4.1.2.2. Sales Data

Sales data are obtained from the sponsor company's ERP system. Sales quantity data are used as a proxy for demand data because the company does not systematically keep track of demand forecasts; and therefore, demand data are inconsistent and unreliable for research purposes. A time series plot of sales quantity can be found in Figure 14.

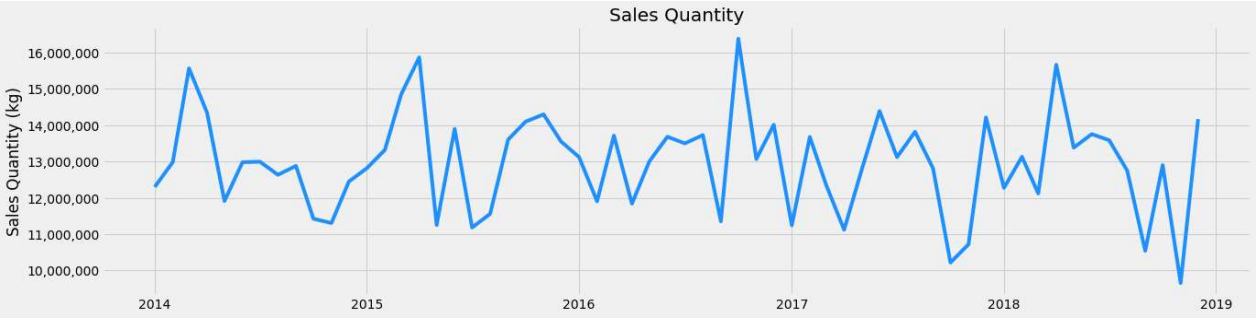


Figure 14. Time Series Plot of Sales Quantity Data

However, sales data are not a perfect substitute for demand data. Apart from the usual reasons in academic literature, such as inventory stockouts (Wecker, 1978), a company-specific reason is that the sponsor company adjusts polystyrene shipments due to plant shutdowns or scheduling issues with freight companies. Since the company's accounting policy is to recognize sales when goods are shipped, these adjustments will result a

difference in demand and sales in the month when the order is placed and the month when the goods are finally shipped.

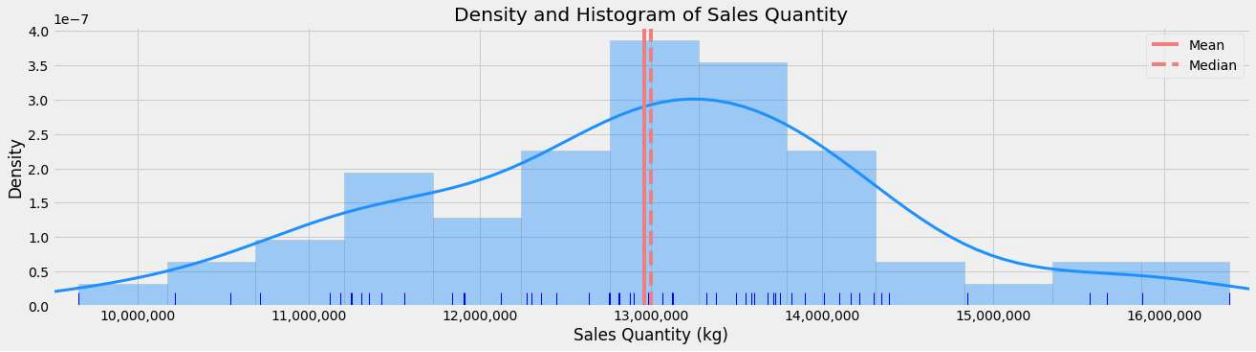


Figure 15. Density Plot and Histogram of Sales Quantity Data

From the density plot and histogram of the sales quantity data in Figure 15, it can be observed that the data are relatively not skewed, and the tails are relatively light compared to the rest of the distribution. This can be checked with the skewness and excess kurtosis measures in Table 5. The measures are close to zero, indicating that the distribution of the sales quantity data is normal. The mean and median are also approximately equal at 12,961,198 and 12,998,505 respectively, which is further evidence that the distribution is symmetrical. This also suggests that although sales are volatile, peaks in demand are counterbalanced by dips in demand.

Table 5. Skewness and Kurtosis Measures of Sales Quantity Data

	Corrected for Bias	Uncorrected for Bias
Skewness	0.01659	0.01617
Excess Kurtosis	0.16549	0.05366

4.2. Results

As discussed in the methodology chapter, the data above are fed into three different models, fine-tuned and evaluated using walk-forward cross-validation against RMSE and MAE as error measures. The sections below describe the predictions obtained from the different models and their relative performances.

For time series plots between Figures 16 to 29, the actual demand is represented by the red line, predictions from the validation dataset are represented by the yellow line and predictions from the test dataset are represented by the blue line. The range of predictions is represented by the lighter shaded area and the darker shaded area represents one standard deviation from the mean.

4.2.1. Simple Average Model

To set a baseline, a forecast using a simple average of actual demand in the previous six months is calculated. The sponsor company uses a similar method to set a baseline forecast. As seen from both Figure 16 and Table 6, this method is not very accurate. In addition, since this model is not cross validated, the error measures in Table 6 should only be used for reference and not for comparison with the other following models.

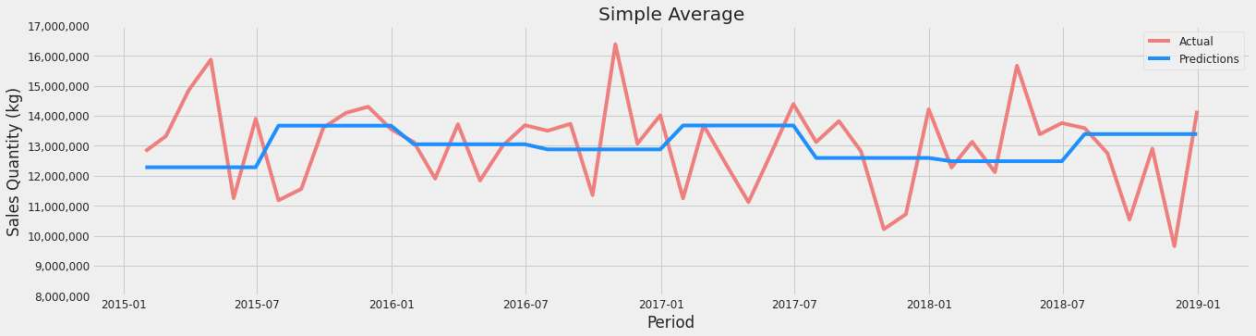


Figure 16. Time Series Plot of Actual and Predicted Sales Quantity using a Simple Half-Year Average

Table 6. Error Measures for the Simple Average Model

RMSE	MAE
1,594,758	1,226,378

4.2.2. ARIMA Model

The `forecast` package in R is used for the modeling of ARIMA models. The ARIMA models generated predictions that did not vary very much from the mean, which is about 13 million. As seen from Figures 17 and 18, the models minimizing for RMSE and MAE resulted in similar predictions that almost form a straight line. More precisely, following the approach described in Section 3.2.4.2, ARIMA models that were either ARIMA (1,0,1) or ARIMA (1,0,2) were obtained. The only difference between the model that minimized for RMSE and the model that minimized for MAE is in the second cross-validation fold. Furthermore, the `nsdiffs()` function returned a value of 0, suggesting that the data has no seasonality.

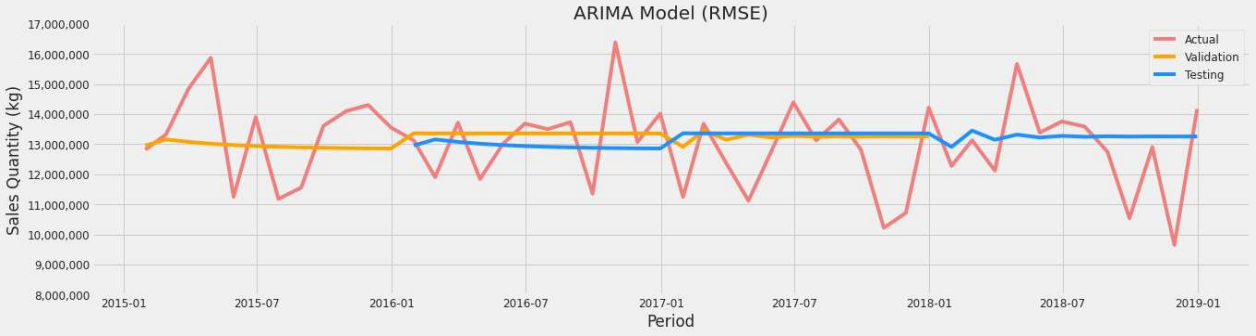


Figure 17. Time Series Plot of Actual and Predicted Sales Quantity for the ARIMA Model, minimizing for RMSE

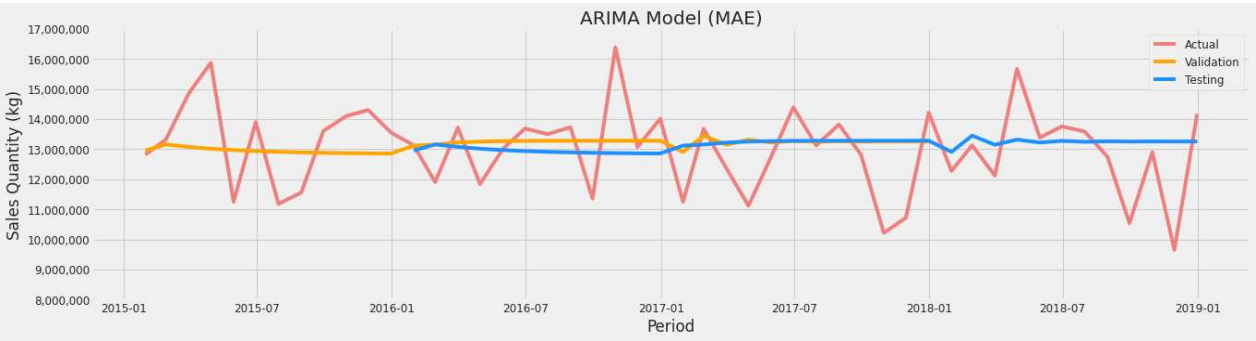


Figure 18. Time Series Plot of Actual and Predicted Sales Quantity for the ARIMA Model, Minimizing for MAE

An interesting observation is that an autoregressive model of order 1, AR (1), was obtained in both versions of the ARIMA model. This is not expected, as Kannegiesser et al. (2009) had pointed out that the demand of chemical commodities is not autoregressive. However, this inconsistency may be because sales quantity was used in lieu of demand quantity. Another reason could be because, unlike the crude-oil related chemical commodities mentioned in Kannegiesser et al., polystyrene is a downstream derivative of styrene; and therefore, is not directly influenced by movements in crude-oil prices. Tables 7 and 8 detail the actual results of the ARIMA models, minimizing for RMSE and MAE respectively.

Table 7. Average Validation and Test Results for the ARIMA Model, Minimizing for RMSE

	RMSE	MAE
Validation	695,777	550,183
Test	743,985	561,700

Table 8. Average Validation and Test Results for the ARIMA Model, Minimizing for MAE

	RMSE	MAE
Validation	693,166	546,662
Test	734,111	555,492

4.2.3. XGBoost Model

The XGBoost models returned a range of predictions that are largely consistent. Figures 19 and 20 show that the model minimizing for RMSE and the model minimizing for MAE do not differ much except for the period from January 2018 to September 2018, in which the latter model predicted a lower demand as compared to the former.

As mentioned in the Methodology chapter (Chapter 3), the news articles' metadata and the related weekly price indices are used to predict demand. As the frequency of news articles is daily, there should be multiple predictions for a month if daily data is used to

predict monthly demand. However, the XGBoost models' predictions look solid and consistent except for the slightly shaded regions of the blue line at the start and around November 2016 in Figure 19. This is evidence that although there are over 15,000 metadata variables from news articles to choose from, the XGBoost models did not use many of them and chose to use common variables present across the training dataset, such as "Month," in their predictions.

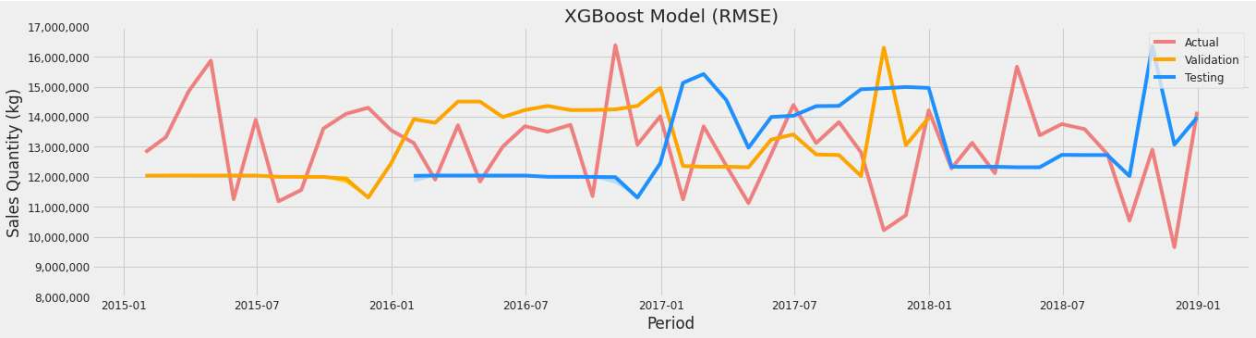


Figure 19. Time Series Plot of Actual and Predicted Sales Quantity for the XGBoost Model, Minimizing for RMSE

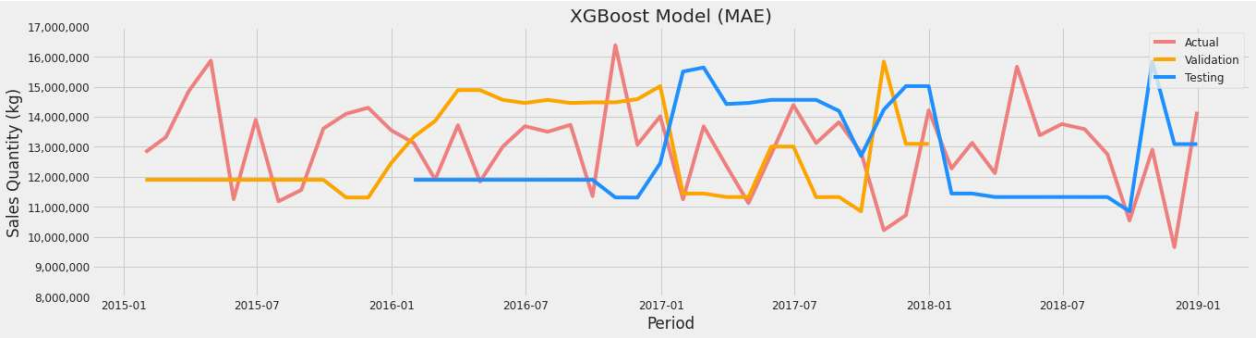


Figure 20. Time Series Plot of Actual and Predicted Sales Quantity for the XGBoost Model, Minimizing for MAE

This is further supported by Figure 21, which shows the feature importance of each feature learned by the ensemble model from the training dataset in one of the three cross-validation folds. The feature importance, or the F Score, is the number of times each feature is split on by each decision tree. From Figure 21, it can be seen that "Month" is one of the most used features to split on. However, as shown by Figure 22, this does not

mean “Month” is the most important feature in a decision tree; it is simply the feature that is used most often across all the decision trees learned by the XGBoost model in that particular cross-validation fold.

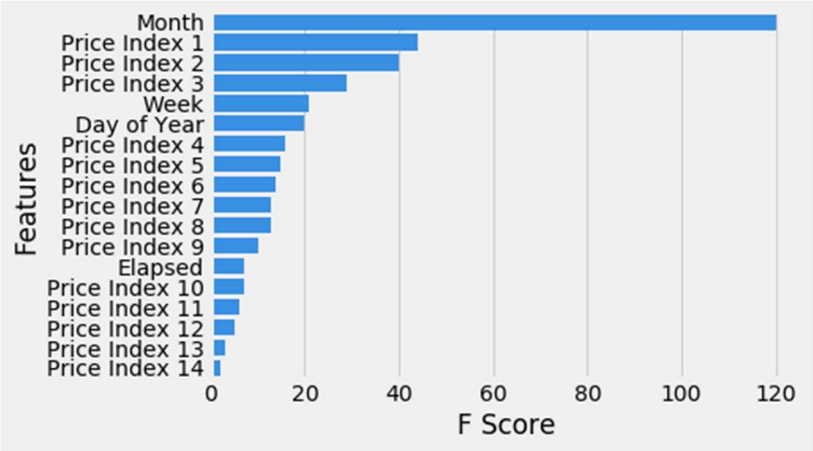


Figure 21. Example of Feature Importance Rank in One of the Three Cross-Validation Folds, in the XGBoost Model Minimizing for RMSE

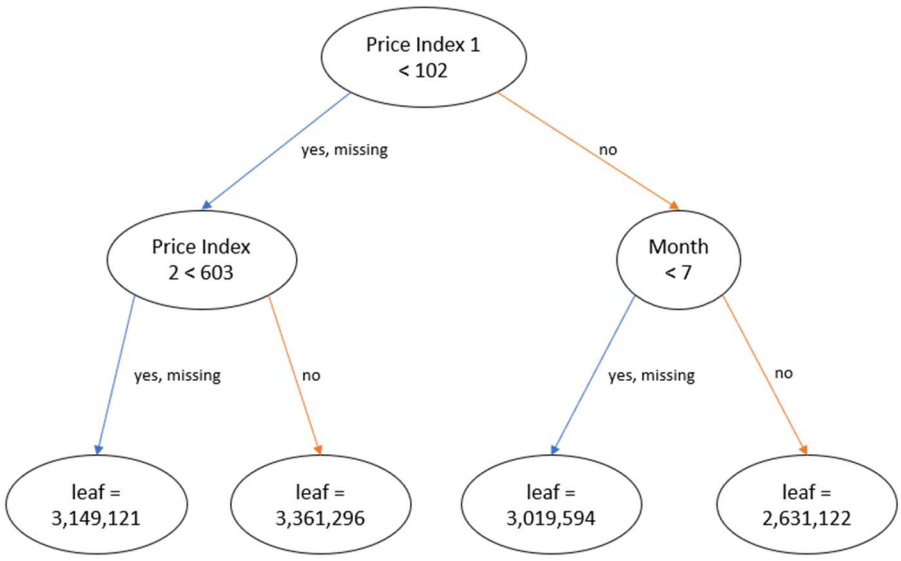


Figure 22. Example of a Decision Tree Learned by the XGBoost Model Minimizing for RMSE

However, the XGBoost models did not give good predictions. As seen from Figures 19 and 20, they predicted the exact same quantity for long periods of time and were not able to

capture the peaks and troughs very well. Both models predicted a rise in demand around January 2017 and November 2018 but were unable to predict the steep decline in demand around the end of 2018. The poor predictions can also be validated quantitatively from the error measures shown in Tables 9 and 10.

Table 9. Average Validation and Test Results for the Best Model, Minimizing for RMSE

	RMSE	MAE
Validation	933,108	733,825
Test	1,022,240	807,572

Table 10. Average Validation and Test Results for the Best Model, Minimizing for MAE

	RMSE	MAE
Validation	1,012,241	837,645
Test	1,133,460	925,622

4.2.4. NEMO

Several variations of NEMO were used, as minor changes were made to the model architecture and parameters were adjusted to fine-tune the model. In general, a model took about 90-120 minutes to train over 6 epochs, depending on the type of GPU used. As a model had to be trained three times across each cross-validation fold, it took between 4.5 to 6 hours to fully evaluate a model.

The performance of the initial working model and the subsequent best model is described in the sections below. In addition, due to the data distribution problem, an alternative model trained over a longer period of two years is also analyzed.

4.2.4.1. Initial Working Model of NEMO

The first working model returned a range of demand forecasts for each month, represented by the shaded regions in Figures 23 and 24. From the same figures, it can be seen that the predictions fluctuate monthly and less smoothed as compared to the ARIMA and XGBoost models.

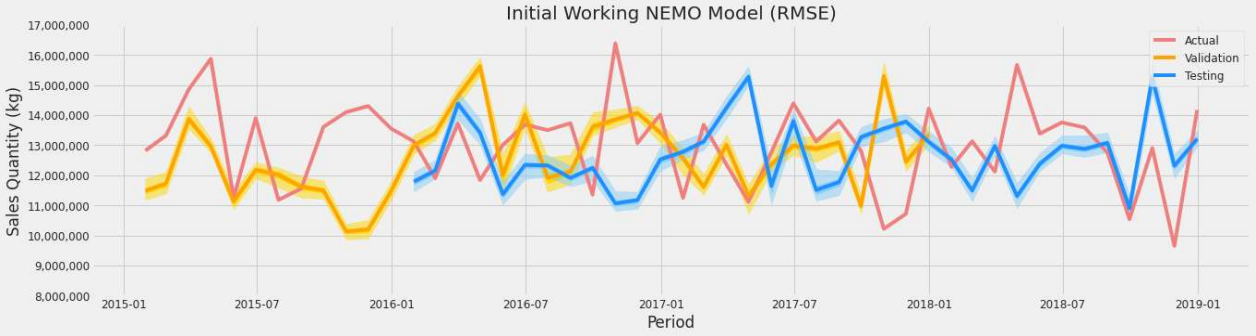


Figure 23. Time Series Plot of Actual and Predicted Sales Quantity for the Initial Working Model of NEMO, Minimizing for RMSE

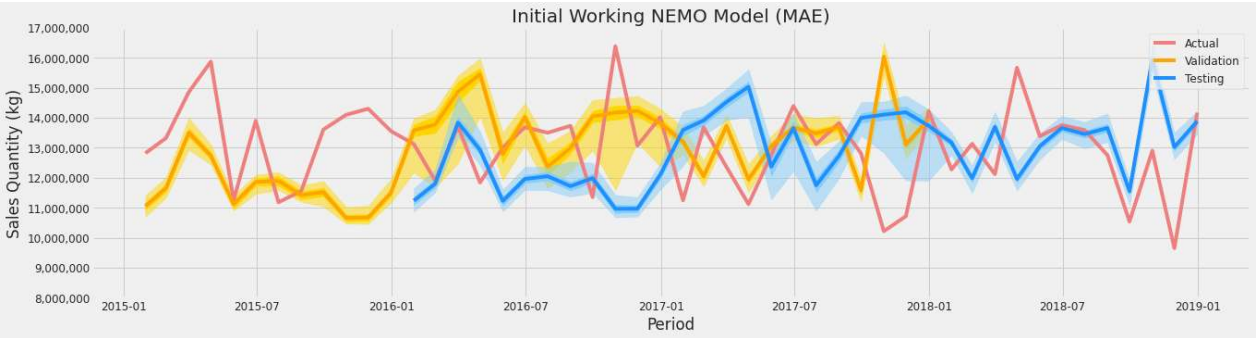


Figure 24. Time Series Plot of Actual and Predicted Sales Quantity for the Initial Working Model of NEMO, Minimizing for MAE

The model that minimized for RMSE and the model that minimized for MAE resulted in similar average demand forecasts, as seen from the blue lines above. However, the latter's predictions deviated more from the actual demand. This is in line with the expectations of the different accuracy measures, as explained in Section 3.2.4.1. It suggests that the model minimizing for MAE penalizes error less, leading to a larger range of forecast values. This can also be verified quantitatively by referring to Tables 11 and 12, where the model that minimized for RMSE gave better results for both RMSE and MAE. However, the difference is very slight, and it is unclear which is a better measure here.

Table 11. Average Validation and Test Results for the Initial Working Model of NEMO, Minimizing for RMSE

	RMSE	MAE
Validation	971,318	771,604
Test	986,850	792,641

Table 12. Average Validation and Test Results for the Initial Working Model of NEMO, Minimizing for MAE

	RMSE	MAE
Validation	981,178	762,297
Test	1,031,187	805,490

4.2.4.2. Best NEMO Model

The best NEMO model uses four more fully connected linear layers with rectified linear activation unit (ReLU) activation functions between the concatenated layer and the final layer. This allowed NEMO to extract more features from the dataset. In addition, various learning rates were tried in an experimental manner to find the best ones. These minor modifications improved the performance of NEMO significantly.

As seen from Figures 25 and 26, the average predictions of the best model, as depicted by the blue lines, are more smoothed than the initial model but still managed to predict some of the peaks and troughs of the actual demand. The range of the predictions, as shown by the shaded regions, is wider than the initial model's. The forecast range of the model minimizing RMSE manages to cover most of the actual demand.

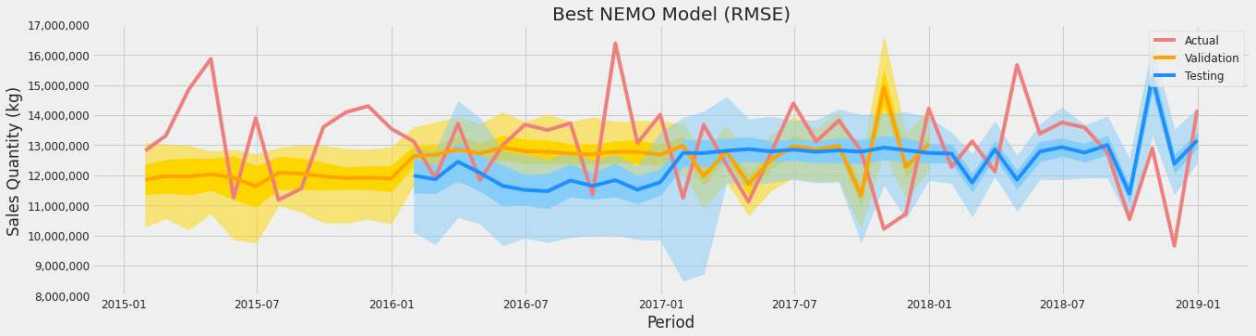


Figure 25. Time Series Plot of Actual and Predicted Sales Quantity for NEMO, Minimizing for RMSE

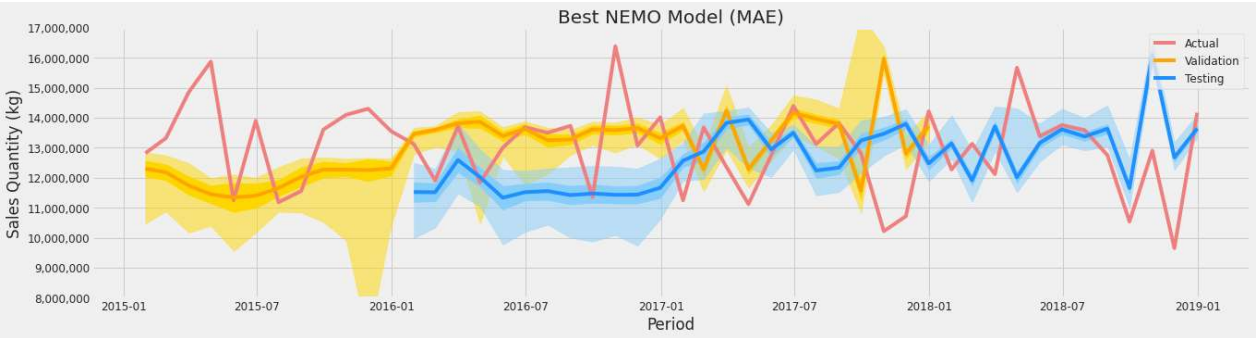
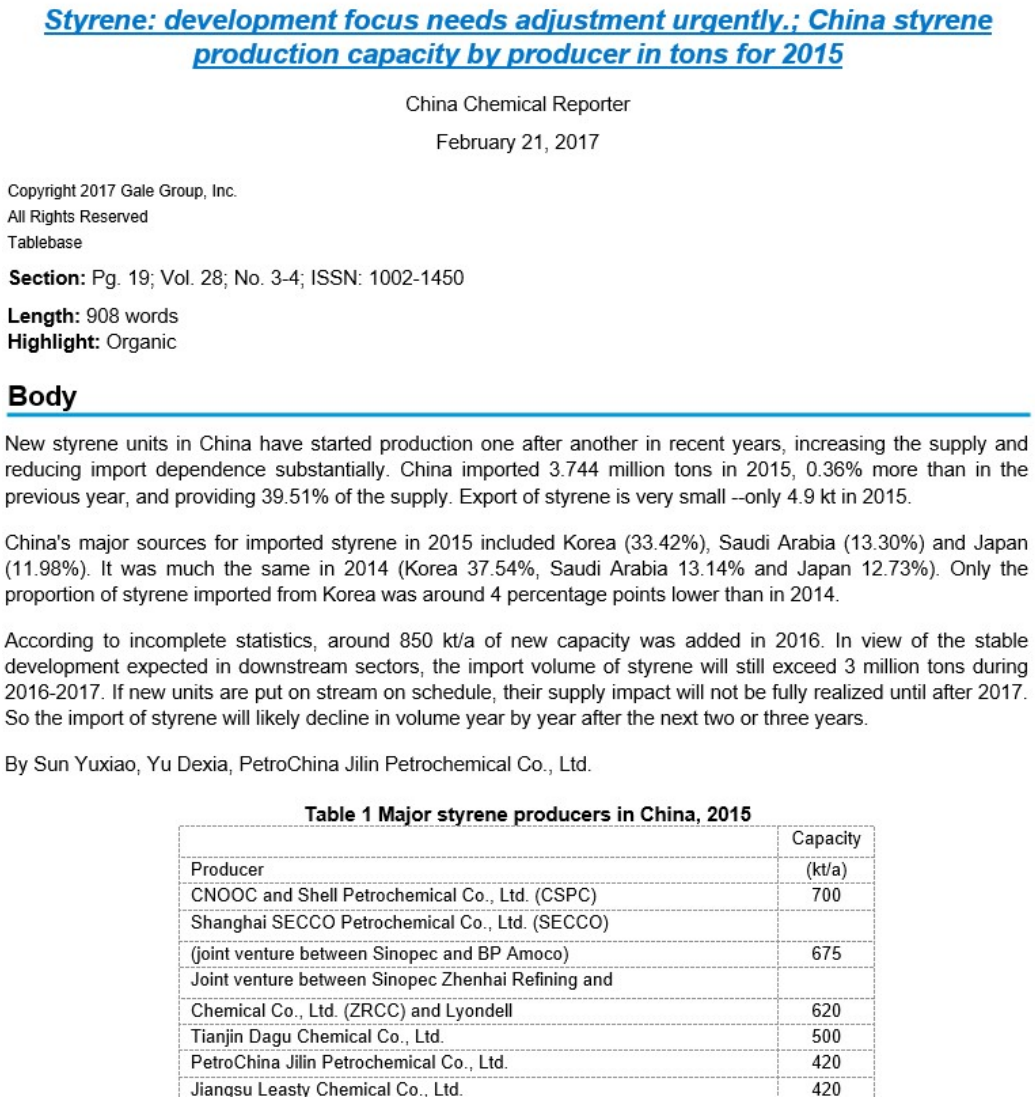


Figure 26. Time Series Plot of Actual and Predicted Sales Quantity for NEMO, Minimizing for MAE

However, the model that minimized for RMSE had two outliers in its predictions. Both predictions were an order of magnitude higher than the average and hence excluded from the results. Looking closer at the data, the documents associated with the predictions are an article on China's ban on imported scrap, and another on China's styrene production capacity. A modified excerpt from the latter article can be seen in Figure 27. Both articles are longer than average, at about 75% quantile. However, the content of the articles is very different. It is hard to determine the reason or the variable that caused the model to predict such high demand. However, the article in Figure 27 has tables about styrene capacity and plants that were planned for construction in China. Styrene monomer is a key raw material of polystyrene, making up over 90% of the bill of materials. One possible

reason could be that having many numbers of a key raw material for a key market threw the model off in its prediction.



Styrene: development focus needs adjustment urgently.; China styrene production capacity by producer in tons for 2015

China Chemical Reporter

February 21, 2017

Section: Pg. 19; Vol. 28; No. 3-4; ISSN: 1002-1450

Length: 908 words
Highlight: Organic

Body

New styrene units in China have started production one after another in recent years, increasing the supply and reducing import dependence substantially. China imported 3.744 million tons in 2015, 0.36% more than in the previous year, and providing 39.51% of the supply. Export of styrene is very small --only 4.9 kt in 2015.

China's major sources for imported styrene in 2015 included Korea (33.42%), Saudi Arabia (13.30%) and Japan (11.98%). It was much the same in 2014 (Korea 37.54%, Saudi Arabia 13.14% and Japan 12.73%). Only the proportion of styrene imported from Korea was around 4 percentage points lower than in 2014.

According to incomplete statistics, around 850 kt/a of new capacity was added in 2016. In view of the stable development expected in downstream sectors, the import volume of styrene will still exceed 3 million tons during 2016-2017. If new units are put on stream on schedule, their supply impact will not be fully realized until after 2017. So the import of styrene will likely decline in volume year by year after the next two or three years.

By Sun Yuxiao, Yu Dexia, PetroChina Jilin Petrochemical Co., Ltd.

Table 1 Major styrene producers in China, 2015

	Capacity
Producer	(kt/a)
CNOOC and Shell Petrochemical Co., Ltd. (CSPC)	700
Shanghai SECCO Petrochemical Co., Ltd. (SECCO)	
(joint venture between Sinopec and BP Amoco)	675
Joint venture between Sinopec Zhenhai Refining and	
Chemical Co., Ltd. (ZRCC) and Lyondell	620
Tianjin Dagu Chemical Co., Ltd.	500
PetroChina Jilin Petrochemical Co., Ltd.	420
Jiangsu Leasty Chemical Co., Ltd.	420

Figure 27. Excerpt from Article that Resulted in an Outlier

One version of NEMO attempted to constrain the range of predictions by using a sigmoid function on the output layer. However, this model was unable to generalize well and thus discarded. As expected, the model that minimized for MAE did not have any issues with outliers and predicted a reasonable range of values for demand. This is because MAE is

robust and not sensitive towards outliers. However, as seen from Tables 13 and 14, the model minimizing for MAE returned higher errors, as measured by RMSE and MAE on the testing set.

Table 13. Average Validation and Test Results for the Best Model, Minimizing for RMSE

	RMSE	MAE
Validation	851,368	689,868
Test	838,229	673,035

Table 14. Average Validation and Test Results for the Best Model, Minimizing for MAE

	RMSE	MAE
Validation	907,563	679,239
Test	958,106	770,273

4.2.4.3. Best NEMO Model over a Longer Training Period

As deep learning models tend to perform better with more data, the same model but trained with two years' worth of data is also evaluated. Correspondingly, the number of cross-validation folds had to be reduced from three to two.

As Figures 28 and 29 show, this version of NEMO returned predictions that are relatively smoothed and not far from the actual demand. This can be validated from Tables 15 and 16, which shows this model has the lowest RMSEs for the testing set out of all the models except for the ARIMA model. However, since this model is evaluated on a different number of cross-validation folds and hence did not have the large spike in demand in October 2016 in its testing data, its results are not directly comparable with the models above.

An interesting observation is that the model minimizing for RMSE returned a higher RMSE and a lower MAE on the testing set as compared to the model minimizing for MAE. This is probably due to the latter model having better predictions in the last quarter of

2018, as seen in Figure 29 where the red line is tracked closely by the blue line and is covered entirely by the blue shaded region. Nevertheless, the difference is very slight at about 2% of the total and can be attributed to randomness.

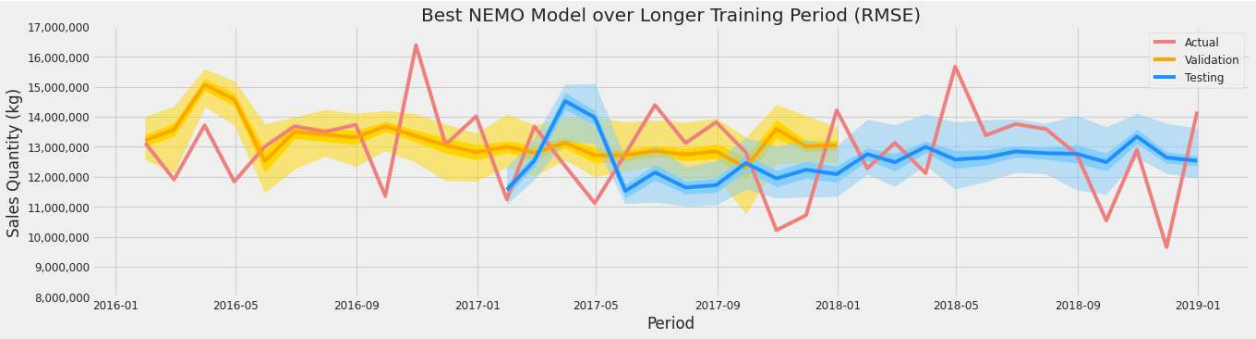


Figure 28. Time Series Plot of Actual and Predicted Sales Quantity for NEMO, Trained for 2 Years, Minimizing for RMSE

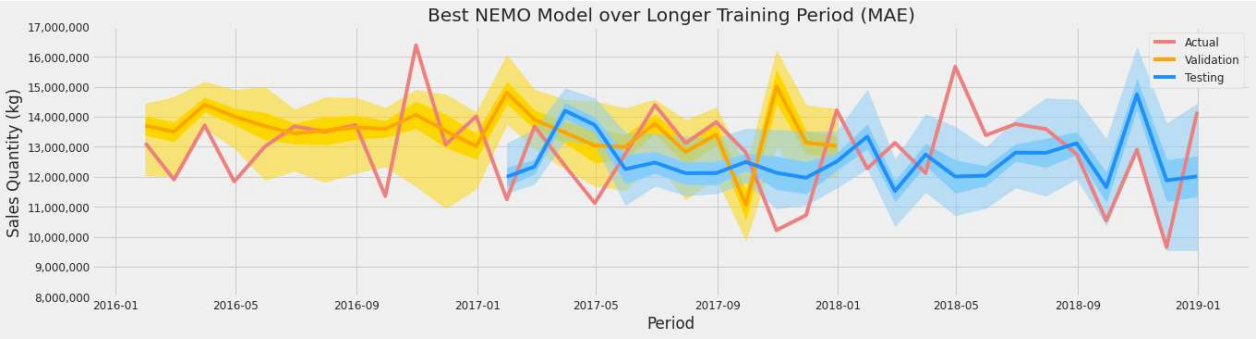


Figure 29. Time Series Plot of Actual and Predicted Sales Quantity for NEMO, Trained for 2 Years, Minimizing for MAE

Table 15. Average Validation and Test Results for NEMO, Trained for 2 Years, Minimizing for RMSE

	RMSE	MAE
Validation	772,232	601,478
Test	826,854	704,918

Table 16. Average Validation and Test Results for NEMO, Trained for 2 Years, Minimizing for MAE

	RMSE	MAE
Validation	842,882	638,143
Test	814,390	720,820

4.2.5. Comparative Analysis

In this section, the performances of the various models, minimizing for RMSE and MAE, are compared and discussed.

4.2.5.1. Comparing across Models

Of the three models, it can be seen from Figures 30 to 33 that the ARIMA model has the lowest error measures, irrespective of minimizing for RMSE or MAE. However, as discussed in Section 4.1.2.2, the distribution of actual sales data is normal and centered around 13,000,000. The ARIMA models' predictions are also centered around 13,000,000 and have standard deviations of less than 200,000 from the mean. Hence, the small amount of error observed is expected. This can be seen in Figures 17 and 18, where the ARIMA models' predictions form almost a straight line and barely track the movements of actual demand; therefore, these models are not ideal for forecasting the erratic demand of polystyrene.

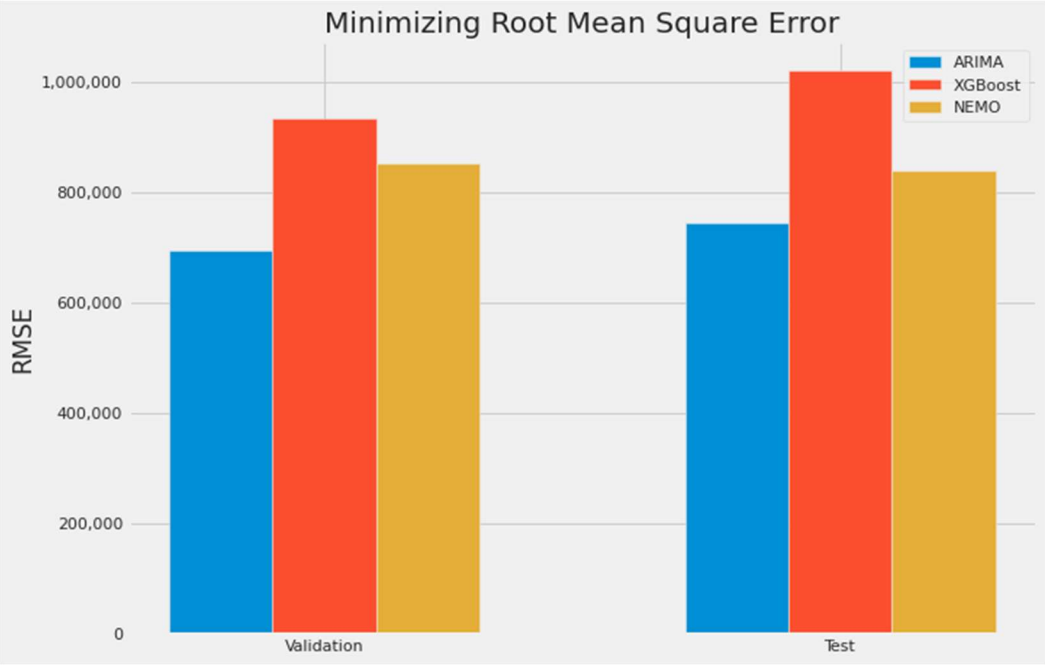


Figure 30. A Comparison of RMSE Scores of Different Models when Minimizing for RMSE

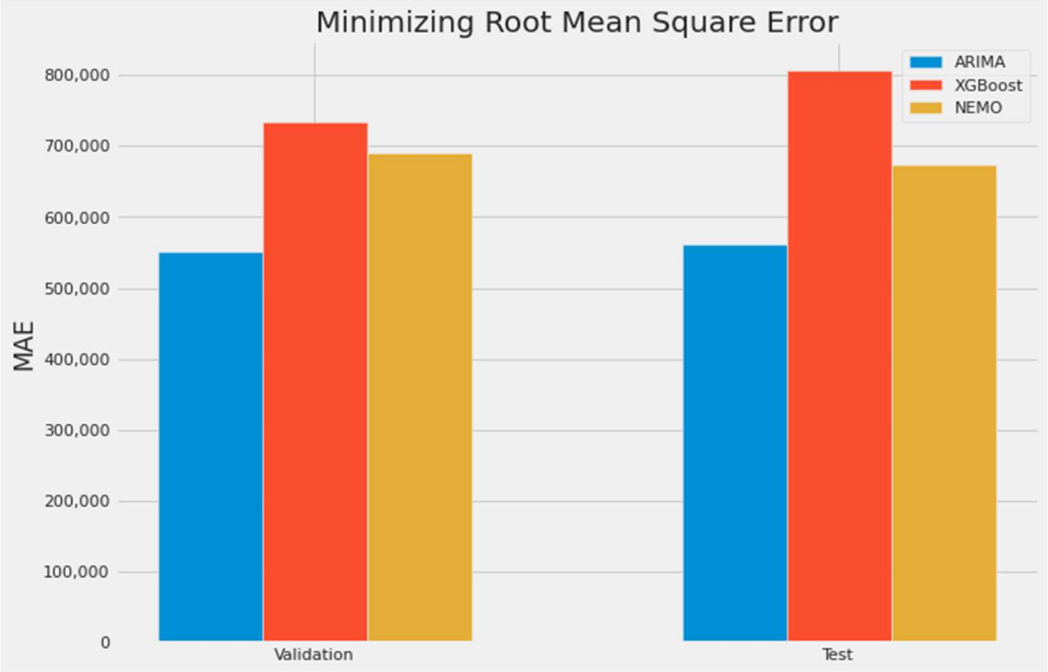


Figure 31. A Comparison of MAE Scores of Different Models when Minimizing for RMSE

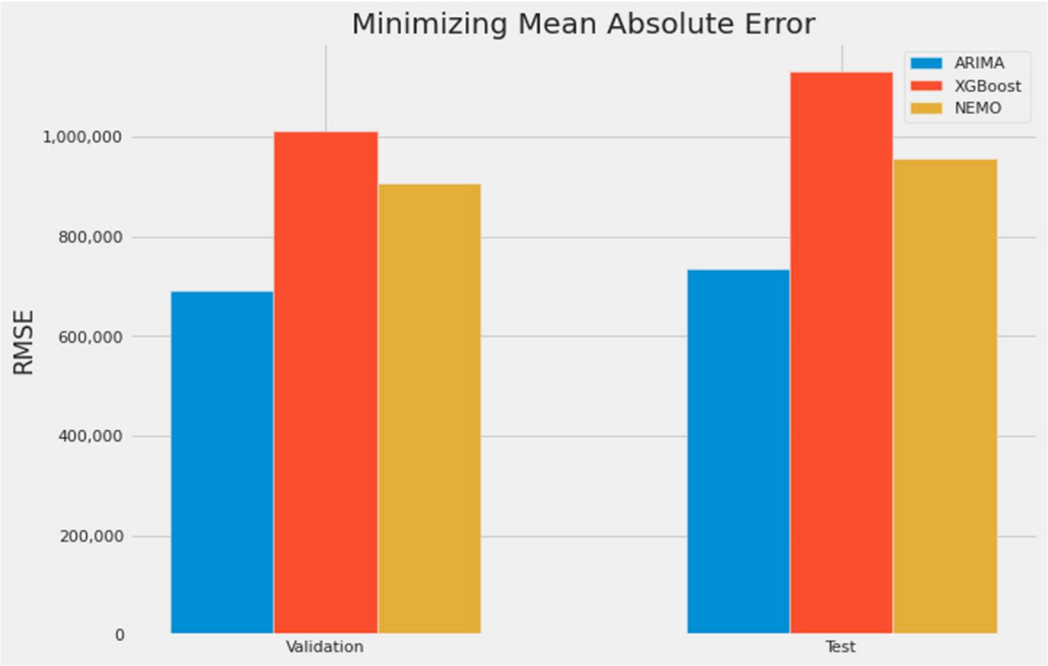


Figure 32. A Comparison of RMSE Scores of Different Models when Minimizing for MAE

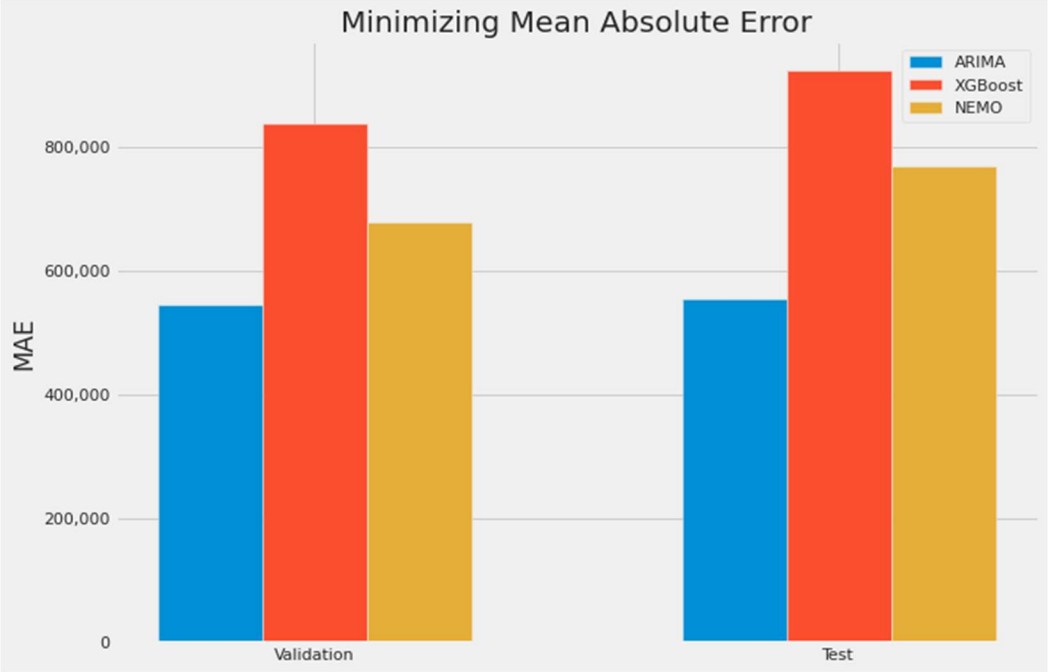


Figure 33. A Comparison of MAE Scores of Different Models when Minimizing for MAE

Between the XGBoost model and NEMO, it can be observed from Figures 30 to 33 that NEMO is superior in both measures, offering approximately 20% better performance when measured on the testing dataset regardless of which measure the model minimizing for. Furthermore, the XGBoost model performed noticeably worse on the testing dataset as compared to the validation dataset, which shows the XGBoost model did not generalize well.

4.2.5.2. Minimizing for RMSE or MAE

From Table 17, it can be observed that models minimizing for RMSE have lower errors than models that minimized for MAE, except for the ARIMA model. This is because the ARIMA model only uses sales data for prediction and since the underlying distribution is not skewed and normal, the predictions are also similar. Conversely, the difference in performance between the models that minimized for RMSE and for MAE, for both the XGBoost and NEMO, suggests that the distributions of the price indices data and textual

data may be uneven, leading to a biased forecast when training a model that minimizes for MAE.

Table 17. A Comparison of the Percentage Differences in Error when Minimizing for MAE over RMSE

		ARIMA	XGBoost	NEMO
Validation	RMSE	0.38%	-7.82%	-6.19%
	MAE	0.64%	-12.39%	1.56%
Test	RMSE	1.35%	-9.81%	-12.51%
	MAE	1.12%	-12.75%	-12.62%

However, comparing both versions of a model is useful because, as explained in Section 3.2.4.1, models that minimize for RMSE and models that minimize for MAE have different purposes. Minimizing for RMSE will result in an unbiased forecast while a model that minimizes for MAE is robust to outliers, as shown by NEMO's results.

4.3. Summary

Regarding the proposed hypotheses, the difference in predictions between the XGBoost model and NEMO suggests that textual documents contain some information that can be used to forecast future demand. However, given the large errors returned by NEMO, the results do not support Hypotheses One and Two, which postulate that news articles can forecast the demand of B2B commodities accurately and NLP techniques improves forecast accuracy and timeliness.

Some of the forecast errors may be because sales data are used in lieu of demand data, which contain irregularities such as unplanned plant shutdowns. Such information is not included in any datasets, which leads to a large forecasting error. Another reason may be due to the distribution of the textual data, as NEMO seemed to perform better in the alternative model, which is trained on bigger dataset. Hence, NEMO's results remain

promising and it would be interesting to test NEMO on actual demand data or on a bigger dataset for verification.

Nevertheless, the results indicate that NEMO is the best for forecasting volatile demand as it offers a middle ground between the three models considered. As seen from Figures 34 and 35, NEMO's predictions are not as smoothed as the ARIMA model, yet NEMO is better in predicting the changes in demand as compared to the XGBoost model. While the predictions are not ideal, NEMO can still be useful in various other contexts, which will be discussed in the following chapter.

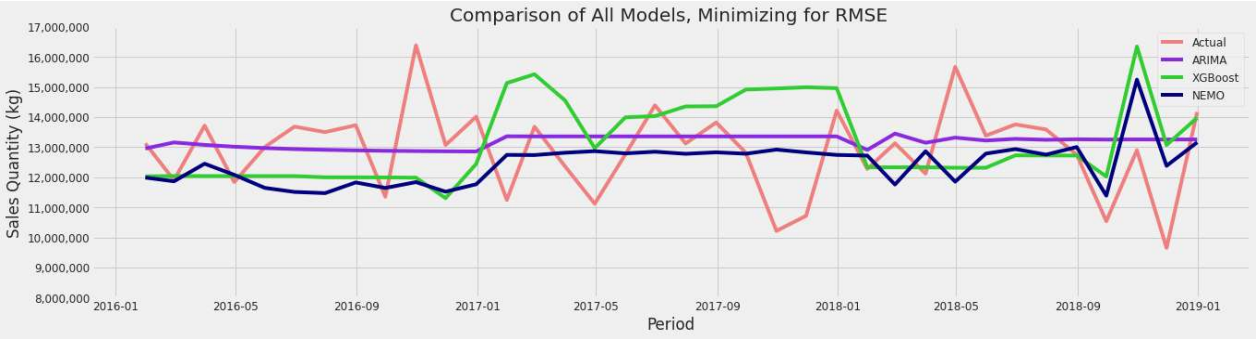


Figure 34. A Comparison of Time Series Plots of Different Models, Minimizing for RMSE

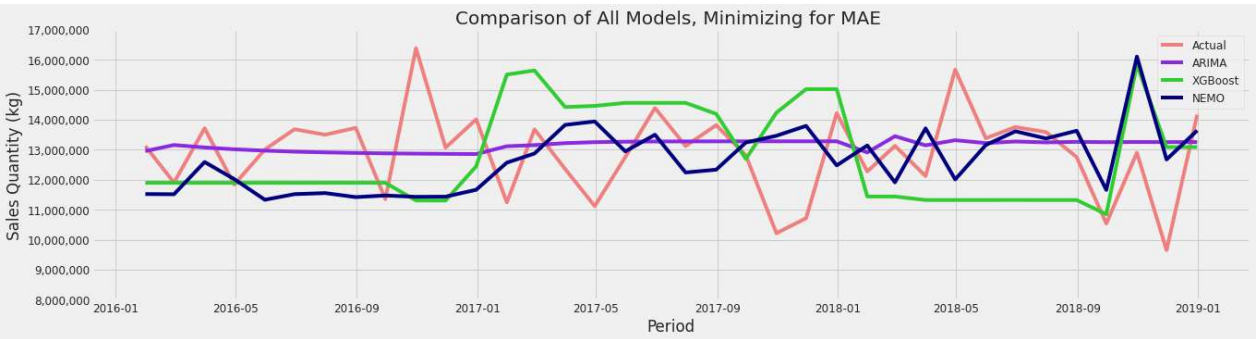


Figure 35. A Comparison of Time Series Plots of Different Models, Minimizing for MAE

5. Discussion

In the last chapter, data and results of various forecasting techniques are explored. In this chapter, implications of the findings will be discussed. This includes practical issues that may be encountered when deploying NEMO in practice. Also, the limitations of NEMO will be touched upon. Lastly, topics for further research will be identified. Broadly, this chapter will contain three main sections.

5.1. Implications

Looking at the results obtained, the sponsor company's current practice of using a simple average of past data, as discussed in Section 4.2.1, combined with expert judgments from experienced salespeople to make forecasting decisions, can be improved even by using a simple ARIMA model. However, as discussed extensively in Section 4.2.5.1, the ARIMA model is unable to predict the peaks and troughs of the underlying commodity. Therefore, NEMO can be used here as a better baseline model and complement the existing forecasting methodology of the sponsor company.

NEMO also has the added advantage of predicting a range of values for any given month, which can be useful to management for determining the upper bounds and lower bounds of demand, and consequently amount of inventory on hand to hold.

Nevertheless, several practical issues must be considered before NEMO can be deployed in actual operations.

5.1.1. Investment Required

The sponsor company currently does not have the correct environment to use NEMO and has to make substantial investments in systems before it is able to do so. Machine learning systems have all the challenges of traditional software development and more, and most non-technology companies are not equipped to handle these challenges. A group of

Google researchers highlighted these challenges as the "technical debt" of machine learning, and warned that such technical debt will compound quickly if not implemented appropriately on a systems level (Sculley et al., 2015). Examples of issues to consider include how to handle ongoing maintenance and how to make adjustments to complex deep learning models. As an analogy, while it is easy to set up a home server to host a personal website, hosting a business website to handle e-commerce comes with its own set of considerations, such as reliability and regulatory concerns. In fact, large technology companies such as Google and Facebook have created their own machine learning platforms so that they can build and deploy machine learning solutions consistently and reliably.

Conversely, a simple model such as an ARIMA model or a logistic regression model can be easily run off Excel using a standard workstation. Therefore, a company should weigh the need for a cutting-edge machine learning model against the personnel and resources needed to deploy such a model.

5.1.2. Training Time Considerations

Even with the correct systems in place, a deep learning model can still take a considerable amount of time to train. For example, NEMO took about 90 minutes to train with a modern GPU with a relatively small dataset. Modern deep learning models have much more complex architectures and use datasets that are hundreds of times larger. Such models require days to be fully trained. For example, BERT took 4 days to be trained on 16 TPU chips, which are Google's custom-made chips developed specifically for the purposes of training neural networks (Devlin et al., 2018). While a language model does not need to be retrained every day, having a long training time in general will have an

impact on operations, especially if there is a limited pool of computing resources and multiple forecasting models for various product lines to be trained.

5.1.3. Textual Data Collection

NEMO requires large amounts of textual data to be collected. Textual data are readily available from news databases or can be scrapped from the web. However, there are legal implications as well as intellectual property and copyright issues surrounding text mining. Scraping web content is a legal gray area, but commercial usage of such content is usually disallowed. Some databases allow text mining, but they charge a hefty premium to do so. Any company considering adopting NEMO must consider such issues.

5.2. Limitations

In addition to the practical issues to be considered above, NEMO has several limitations, which are discussed below.

5.2.1. Black Box Model

NEMO is a deep learning model and a big limitation of such models is they can be black box models. Consequently, deep learning models are challenging to understand and hard to debug. For example, the outliers predicted by NEMO, detailed in Section 4.2.4.2, were challenging to explain as there are thousands of variables to consider. Furthermore, there is usually a trade-off between performance and interpretability in such models, where accurate models are often exponentially more complex and interpretable models are usually less accurate (Johansson, Sönströd, Norinder, & Boström, 2011).

There is a lot of ongoing research focused on improving interpretability of black box models, with model-agnostic methods like local interpretable model-agnostic explanations (LIME) (Ribeiro, Singh, & Guestrin, 2016) and Shapley additive explanations (SHAP) (Lundberg & Lee, 2017). These methods uses a property called local

explainability and try to understand model decisions for a particular data point, as opposed to trying to understand the model as a whole (Molnar, 2019).

5.2.2. Disinformation and Adversarial Attacks

Another significant limitation of NEMO is its susceptibility to disinformation. Apart from the usual method of hiring humans to write "fake" news or reviews, NLP techniques have advanced to the point where artificially generated text is indistinguishable from those written by a human being. For example, OpenAI's GPT-2 model alarmed its developers so much that they decided to release the code for GPT-2 in stages, due to their concern that GPT-2 will be used for malicious purposes at scale (Radford et al., 2019).

In business, companies may intentionally generate fake news to limit competition or artificially inflate demand. Such deliberate acts of disinformation are known as adversarial attacks, which seek to confound machine learning systems purposely and is an emerging field of research as well. Zhou, Guan, Moorthy Bhat, and Hsu (2019) showed NLP models can be vulnerable to such attacks, but their proposed solution of a crowdsourced knowledge graph may not be practical to implement concurrently with NEMO.

5.2.3. Novel Vocabulary

As a machine learning model, NEMO is able to learn only from past events to predict future demand. In other words, it will not be able to fully understand events that have not happened before, as it has not learned to associate new vocabulary from these events with demand. Furthermore, the language model's vocabulary is limited a certain size to help with performance. Words that are not frequently encountered during training are labeled as "out-of-vocabulary" and disregarded. Therefore, as an example, prior to the 2008

financial crisis, the model will not be able to associate words like “Lehman shock” with a shape decline in demand.

However, this is not a significant limitation as NEMO uses word embeddings, which understand word context. As new vocabulary will appear in the context with other regular words, the model should still be able to understand the general context of the text. Using the previous example, although the model cannot interpret the significance of “Lehman shock” on its first encounter, the model will be able to correlate the context of a “financial crisis” with a drop in demand.

5.3. Further Research

Given the timeliness of this thesis’s research focus, there are many unexplored areas for further research, which can be broadly categorized into data-related and model-related sections.

5.3.1. Data

As with any machine learning model, NEMO’s performance is highly dependent on its data. Data can be varied in many ways to test the performance and extend the applications of NEMO.

Because all companies use the written language in one form or another, the model can analyze a wide range of textual data for prediction. Examples of such data may include emails, written contracts or product reviews. For example, as discussed in Section 1.4, the input data from a B2C company can be quite different from a B2B company in terms of composition and semantics, and may include opinionated data such as customer feedback. It would be interesting to use NEMO with a wider range of data from other industries.

Taking the idea of using other types of data further, NEMO can even be extended to analyze other forms of data, such as audio data. This will unlock many other sources of

information. In the case of audio data, new sources of information could now include news clips and interviews. The only additional step required to do so is to convert audio into text, which can be accomplished automatically with a high accuracy rate today.

5.3.1.1. Bias

The way data are represented can be biased and this is especially true for textual data. Language is inherently biased. For example, research has shown that “He is a doctor” has a higher conditional likelihood than “She is a doctor” in a language model (Lu, Mardziel, Wu, & Amancharla, 2018). Crawford (2017) categorized bias in terms of allocation and representation bias. Allocation bias occurs when a model performs better on data that have greater frequency. A representation bias is when a biased concept is captured by a model due to the way the data are represented such as the gender bias in the previous example.

Using a biased dataset can have far reaching implications, as highlighted by Angwin, Larson, Mattu, and Kirchner (2016) in an ProPublica article which exposed how proprietary algorithms used in U.S. courts are biased against African-Americans. Similarly, as this research is carried out using the English language, this makes the data biased towards English-speaking countries; and since the model is a black box model, such subtle biases are hard to uncover. Therefore, as discussed in Section 5.2.1, further research into interpretability of black box models is important and much needed.

5.3.2. Model

NEMO is developed as a proof-of-concept for a very specific problem, so there are many areas for further development and improvement.

5.3.2.1. More Tricks and Tuning

As mentioned in Section 4.2.4, NEMO's performance improved significantly with minor adjustments as simple as varying the learning rate. Currently, tuning deep learning models is more of an art than a science, and there are many other hyperparameters to change and "tricks" that can be tried to further improve the model in terms of accuracy and other performance measures like training time.

Many papers document the different "bags of tricks" that deep learning practitioners and researchers found to improve the performance of their models. For example, He et al. (2019) detailed the usage of 16-bit floating point precision over the standard 32-bit, which resulted in a reduction in the time needed to train their models by 2 to 3 times. Similarly, NEMO can be further improved by implementing some of these "tricks."

In addition, much like the Bayesian optimization library used by the XGBoost model, hyperparameter optimization frameworks such as Optuna for PyTorch models have been developed recently to automate the search for best hyperparameters in deep learning models. Adopting the use of such frameworks will reduce the effort and guesswork needed to find the best hyperparameters for NEMO.

5.3.2.2. Other Neural Networks Variants

NEMO has a relatively simple architecture and uses LSTMs. Modern deep learning networks can be up to thousands of layers deep and use recently developed concepts like self-attention, which allows a neural network to figure out which inputs to pay more attention to. Hence, adding more layers to NEMO and updating some components of the network will probably result in better performance.

Furthermore, NEMO is not explicitly designed to handle time series data. RNNs are particularly well-suited for time series prediction as they can maintain an internal state

which retains the information they have encountered so far. Other promising methods to consider include using convolutional neural networks (CNNs) to extract important features (Pang, Yin, Zhang, & Zhao, 2017) or a combination of CNNs and RNNs (Cirstea, Micu, Muresan, Guo, & Yang, 2018) to handle time series data. All these variants can be used in place of the linear layers after concatenation in NEMO and should improve the model's predictions.

5.3.2.3. Online Learning

NEMO assumes all data are available *a priori* for training, which may not be realistic for real-world applications where data often arrive in streams. An interesting extension for NEMO is to examine the feasibility of adapting the model for online learning. Online learning refers to the machine learning method in which the latest data available are used immediately to update a model's best prediction. However, online learning is substantially more difficult to achieve for deep learning models, as they will experience convergence issues such as vanishing gradients if online learning is applied directly on them. Sahoo, Pham, Lu, and Hoi (2018) have proposed hedge backpropagation to allow online learning to be used with deep learning models. Modifying NEMO to use hedge backpropagation will enable it to make faster predictions without having to retrain the entire model.

5.3.2.4. Multi-step Forecasting

In multi-step forecasting, two or more future time steps are predicted simultaneously, which can be accomplished through various methods. One method is to frame NEMO as a multi-step forecasting model using a sliding window and use RNNs to predict the future demand for the next few steps. Adapting NEMO to predict multiple time steps will be useful for long term planning purposes, such as planning the inventory levels for the next

three months. However, multi-step forecasting also comes with additional complexities, such as error aggregation and a higher level of uncertainty (Bontempi, Taieb, & Le Borgne, 2012).

6. Conclusion

In this thesis, NLP techniques are applied to textual documents in an attempt to forecast the demand of B2B companies selling commodities in long supply chains. Unfortunately, the results do not lend support to NLP techniques being able to predict demand of such commodities accurately. However, this may be due to issues with the underlying dataset. Nevertheless, given that one of the purposes of this thesis is to establish an NLP-based forecasting methodology, NEMO's results are still promising and suggest such that a forecasting method remains viable, but much remains to be done before NEMO can be deployed in day-to-day operations. In its current form, NEMO can be used alongside other forecasting models and provide invaluable information about upcoming volatility in demand.

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine Bias. *ProPublica*. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bengfort, B., Bilbro, R., & Ojeda, T. (2018). *Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning*: O'Reilly Media.

Beverland, M., Lindgreen, A., Napoli, J., Kotler, P., & Pfoertsch, W. (2007). Being Known or Being One of Many: The Need For Brand Management for Business-To-Business (B2B) Companies. *Journal of business & industrial marketing*.

Bontempi, G., Taieb, S. B., & Le Borgne, Y.-A. (2012). *Machine Learning Strategies for Time Series Forecasting*. Paper presented at the European Business Intelligence Summer School.

Boone, T., Boylan, J. E., Fildes, R., Ganeshan, R., & Sanders, N. (2019). Perspectives on Supply Chain Forecasting. *International Journal of Forecasting, 35*(1), 121-127. doi:https://doi.org/10.1016/j.ijforecast.2018.11.002

Caplice, C., & Sheffi, Y. (2006a). Demand Forecasting I: Time Series Analysis. In *ESD.260J Logistics Systems. Fall 2006*. Massachusetts Institute of Technology: MIT OpenCourseWare.

Caplice, C., & Sheffi, Y. (2006b). Demand Forecasting II: Causal Analysis. In *ESD.260J Logistics Systems. Fall 2006*. Massachusetts Institute of Technology: MIT OpenCourseWare.

Carbonneau, R., Laframboise, K., & Vahidov, R. (2008). Application of Machine Learning Techniques for Supply Chain Demand Forecasting. *European Journal of Operational Research, 184*(3), 1140-1154. doi:https://doi.org/10.1016/j.ejor.2006.12.004

Chai, T., & Draxler, R. R. (2014). Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? – Arguments Against Avoiding RMSE in the Literature. *Geoscientific model development, 7*(3), 1247-1250.

Chase, C. (2013). *Demand-Driven Forecasting: A Structured Approach to Forecasting*: Hoboken, New Jersey : John Wiley & Sons, Inc., [2013] Second edition.

Chen, F., Drezner, Z., Ryan, J. K., & Simchi-Levi, D. (2000). Quantifying the Bullwhip Effect in a Simple Supply Chain: The Impact of Forecasting, Lead Times, and Information. *Management science, 46*(3), 436-443.

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. Paper presented at the Proceedings of the 22nd ACM SIGKDD International Conference

on Knowledge Discovery and Data Mining, San Francisco, California, USA. https://doi.org/10.1145/2939672.2939785

Cheng, I.-H., & Xiong, W. (2014). Financialization of Commodity Markets. *Annual Review of Financial Economics, 6*(1), 419-441.

Chomsky, N. (1957). *Syntactic Structures*: Mouton Publishers, The Hague.

Cirstea, R.-G., Micu, D.-V., Muresan, G.-M., Guo, C., & Yang, B. (2018). *Correlated Time Series Forecasting using Multi-Task Deep Neural Networks*. Paper presented at the Proceedings of the 27th ACM International Conference on Information and Knowledge Management, Torino, Italy. https://doi.org/10.1145/3269206.3269310

Crawford, K. (2017). The Trouble with Bias. *Neural Information Processing Systems (NIPS'17)* [Keynote].

Crone, S. F., Hibon, M., & Nikolopoulos, K. (2011). Advances in Forecasting with Neural Networks? Empirical Evidence from the NN3 Competition on Time Series Prediction. *International Journal of Forecasting, 27*(3), 635-660. doi:https://doi.org/10.1016/j.ijforecast.2011.04.001

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805v2*.

Eisenstein, J. (2019). *Introduction to Natural Language Processing*: MIT Press.

Fan, Z.-P., Che, Y.-J., & Chen, Z.-Y. (2017). Product Sales Forecasting Using Online Reviews and Historical Sales Data: A Method Combining the Bass Model and Sentiment Analysis. *Journal of Business Research, 74*, 90-100. doi:https://doi.org/10.1016/j.jbusres.2017.01.010

Giunipero, L. C., & Aly Eltantawy, R. (2004). Securing the Upstream Supply Chain: A Risk Management Approach. *International Journal of Physical Distribution & Logistics Management, 34*(9), 698-713. doi:10.1108/09600030410567478

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*: MIT Press.

Green, K. C., & Armstrong, J. S. (2010). Selection Tree for Forecasting Methods. In selection_tree.pdf (Ed.). Forecastingprinciples.com.

Green, K. C., & Armstrong, J. S. (2012). Demand Forecasting: Evidence-Based Methods. *Available at SSRN 3063308*.

Harvey, A. C. (1990). ARIMA Models. In J. Eatwell, M. Milgate, & P. Newman (Eds.), *Time Series and Statistics* (pp. 22-24). London: Palgrave Macmillan UK.

He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., & Li, M. (2019). *Bag of Tricks for Image Classification with Convolutional Neural Networks*. Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Howard, J., & Gugger, S. (2018). fastai Library: GitHub. Retrieved from https://github.com/fastai/fastai

Howard, J., & Ruder, S. (2018). Universal Language Model Fine-Tuning for Text Classification. *arXiv preprint arXiv:1801.06146*.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice* (2nd ed.): OTexts.

Iankova, S., Davies, I., Archer-Brown, C., Marder, B., & Yau, A. (2019). A Comparison of Social Media Marketing Between B2B, B2C and Mixed Business Models. *Industrial Marketing Management, 81*, 169-179. doi:https://doi.org/10.1016/j.indmarman.2018.01.001

Johansson, U., Sönströd, C., Norinder, U., & Boström, H. (2011). Trade-off between Accuracy and Interpretability for Predictive *in silico* Modeling. *Future Medicinal Chemistry, 3*(6), 647-663. doi:10.4155/fmc.11.23

Johnson, O. (2018). *The Price Reporters : A Guide to PRAs and Commodity Benchmarks*: Abingdon, Oxon ; New York, NY : Routledge, an imprint of the Taylor & Francis Group, 2018.

Kalyani, J., Bharathi, H. N., & Jyothi, R. (2016). Stock Trend Prediction Using News Sentiment Analysis. *International Journal of Computer Science and Information Technology, 8*, 67-76. doi:10.5121/ijcsit.2016.8306

Kannegiesser, M., Günther, H.-O., van Beek, P., Grunow, M., & Habla, C. (2009). Value Chain Management for Commodities: A Case Study from the Chemical Industry. *OR Spectrum, 31*(1), 63-93. doi:10.1007/s00291-008-0124-9

Klein, A., Riekert, M., Kirilov, L., & Leukel, J. (2018). Increasing the Explanatory Power of Investor Sentiment Analysis for Commodities in Online Media. *Lecture Notes in Business Information Processing, 320*, 321-332. doi:10.1007/978-3-319-93931-5_23

Lackman, C. L. (2007). Forecasting Sales for a B2B Product Category: Case of Auto Component Product. *The Journal of Business & Industrial Marketing, 22*(4), 228-235. doi:http://dx.doi.org/10.1108/08858620710754496

Lane, H., Hapke, H., & Howard, C. (2019). *Natural Language Processing in Action: Understanding, Analyzing, and Generating Text with Python*: Manning Publications Company.

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. doi:10.1111/poms.12737

Li, J., Xu, Z., Yu, L., & Tang, L. (2016). Forecasting Oil Price Trends with Sentiment of Online News Articles. *Procedia Computer Science, 91*, 1081-1087. doi:https://doi.org/10.1016/j.procs.2016.07.157

Li, X., Shang, W., & Wang, S. (2018). Text-Based Crude Oil Price Forecasting: A Deep Learning Approach. *International Journal of Forecasting*. doi:https://doi.org/10.1016/j.ijforecast.2018.07.006

Liu, B. (2015). *Opinions, Sentiment, and Emotion in Text*: Cambridge University Press.

Liu, B., & Zhang, L. (2012). A Survey of Opinion Mining and Sentiment Analysis. In C. C. Aggarwal & C. Zhai (Eds.), *Mining Text Data* (pp. 415-463). Boston, MA: Springer US.

Lu, K., Mardziel, P., Wu, F., & Amancharla, P. (2018). *Gender Bias in Neural Natural Language Processing*.

Lundberg, S. M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. Paper presented at the Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The M4 Competition: Results, Findings, Conclusion and Way Forward. *International Journal of Forecasting, 34*(4), 802-808. doi:https://doi.org/10.1016/j.ijforecast.2018.06.001

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality.

MIT Technological Review. (2015). King - Man + Woman = Queen: The Marvelous Mathematics of Computational Linguistics. *Emerging Technology from the arXiv*. Retrieved from https://www.technologyreview.com/s/541356/king-man-woman-queen-the-marvelous-mathematics-of-computational-linguistics/

Mitrea, C. A., Lee, C. K. M., & Wu, Z. (2009). A Comparison between Neural Networks and Traditional Forecasting Methods: A Case Study. *International Journal of Engineering Business Management, 1*, 11. doi:10.5772/6777

Molnar, C. (2019). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*.

Nerlove, M., & Diebold, F. X. (1990). Autoregressive and Moving-average Time-series Processes. In J. Eatwell, M. Milgate, & P. Newman (Eds.), *Time Series and Statistics* (pp. 25-35). London: Palgrave Macmillan UK.

Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2014). *Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks*. Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Pai, P., & Liu, C. (2018). Predicting Vehicle Sales by Sentiment Analysis of Twitter Data and Stock Market Values. *IEEE Access, 6*, 57655-57662. doi:10.1109/ACCESS.2018.2873730

Pang, N., Yin, F., Zhang, X., & Zhao, X. (2017). *A Robust Approach for Multivariate Time Series Forecasting*. Paper presented at the Proceedings of the Eighth International Symposium on Information and Communication Technology, Nha Trang City, Viet Nam. https://doi.org/10.1145/3155133.3155172

Pennington, J., Socher, R., & Manning, C. (2014). *GloVe: Global Vectors For Word Representation*. Paper presented at the Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP).

Radford, A., Wu, J., Amodei, D., Amodei, D., Clark, J., Brundage, M., & Sutskever, I. (2019). Better Language Models and Their Implications. Retrieved from https://openai.com/blog/better-language-models/

Ravi, K., & Ravi, V. (2015). A Survey on Opinion Mining and Sentiment Analysis - Tasks, Approaches and Applications. *Knowledge-Based Systems, 89*, 14-46. doi:https://doi.org/10.1016/j.knosys.2015.06.015

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. Paper presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA. https://doi.org/10.1145/2939672.2939778

Ruder, S. (2019). *Neural Transfer Learning for Natural Language Processing*. National University Of Ireland, Galway,

Sahoo, D., Pham, Q., Lu, J., & Hoi, S. C. H. (2018). *Online Deep Learning: Learning Deep Neural Networks on the Fly*. Paper presented at the Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., . . . Dennison, D. (2015). *Hidden Technical Debt in Machine Learning Systems*. Paper presented at the Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, Montreal, Canada.

Shynkevich, Y., McGinnity, T. M., Coleman, S., & Belatreche, A. (2015, 12-17 July 2015). *Stock Price Prediction Based on Stock-Specific And Sub-Industry-Specific News Articles*. Paper presented at the 2015 International Joint Conference on Neural Networks (IJCNN).

Smales, L. A. (2014). News Sentiment in the Gold Futures Market. *Journal of Banking & Finance, 49*, 275-286. doi:https://doi.org/10.1016/j.jbankfin.2014.09.006

Smith, L. N. (2018). A Disciplined Approach to Neural Network Hyper-Parameters: Part 1 - Learning Rate, Batch Size, Momentum, and Weight Decay. *arXiv preprint arXiv:1803.09820*.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). *Practical Bayesian Optimization of Machine Learning Algorithms*. Paper presented at the Advances in Neural Information Processing Systems.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification? *arXiv preprint arXiv:1905.05583*.

Swain, A. K., & Cao, R. Q. (2013). *Exploring The Impact Of Social Media On Supply Chain Performance: A Sentiment Analysis*. Paper presented at the Decision Sciences Institute Annual Meeting Proceedings.

Swain, A. K., & Cao, R. Q. (2017). Using Sentiment Analysis to Improve Supply Chain Intelligence. *Information Systems Frontiers*, 1-16. doi:10.1007/s10796-017-9762-2

Syntetos, A. A., Babai, Z., Boylan, J. E., Kolassa, S., & Nikolopoulos, K. (2016). Supply Chain Forecasting: Theory, Practice, Their Gap and the Future. *European Journal of Operational Research, 252*(1), 1-26. doi:https://doi.org/10.1016/j.ejor.2015.11.010

Tremblay, E. (2019). Merging Image + Tabular + Text Data in Single Neural Network with fastai. *PetFinder.my Adoption Prediction*. Retrieved from https://www.kaggle.com/c/petfinder-adoption-prediction/discussion/89491

Varma, S., & Simon, R. (2006). Bias in Error Estimation when Using Cross-Validation for Model Selection. *BMC Bioinformatics, 7*(1), 91. doi:10.1186/1471-2105-7-91

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). *Attention is All You Need*. Paper presented at the Advances in Neural Information Processing Systems.

Wecker, W. E. (1978). Predicting Demand from Sales Data in the Presence of Stockouts. *Management science, 24*(10), 1043-1054.

Weller, M., & Crone, S. F. (2012). *Supply Chain Forecasting: Best Practices & Benchmarking Study*: Lancaster Centre for Forecasting.

What Makes Something a Commodity? (2017). *The Economist Explains*. Retrieved from https://www.economist.com/the-economist-explains/2017/01/03/what-makes-something-a-commodity

Wood, L. C., Reiners, T., & Srivastava, H. S. (2013). Expanding Sales and Operations Planning Using Sentiment Analysis: Demand and Sales Clarity from Social Media.

Wood, L. C., Reiners, T., & Srivastava, H. S. (2014). Sentiment Analysis in Supply Chain Management. In W. John (Ed.), *Encyclopedia of Business Analytics and Optimization* (pp. 2147-2158). Hershey, PA, USA: IGI Global.

Wood, L. C., Reiners, T., & Srivastava, H. S. (2015). Exploring Sentiment Analysis to Improve Supply Chain Decisions. *Available at SSRN 2665482*.

Wood, L. C., Reiners, T., & Srivastava, H. S. (2016). Think Exogenous to Excel: Alternative Supply Chain Data to Improve Transparency and Decisions. *International Journal of Logistics*. doi:https://doi.org/10.1080/13675567.2016.1267126

Xu, X., Qi, Y., & Hua, Z. (2010). Forecasting Demand of Commodities After Natural Disasters. *Expert Systems with Applications, 37*(6), 4313-4317. doi:https://doi.org/10.1016/j.eswa.2009.11.069

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv preprint arXiv:1906.08237*.

Zhou, Z., Guan, H., Moorthy Bhat, M., & Hsu, J. (2019). Fake News Detection via NLP is Vulnerable to Adversarial Attacks. *Proceedings of the 11th International Conference on Agents and Artificial Intelligence*. doi:10.5220/0007566307940800