Unstructured Data in Digital Marketing and Supply Chain Management

Xingyi Li

University College London

School of Management

Submitted to University College London (UCL) in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy.

Thesis submission date: Jul 2024

Declaration

I, Xingyi Li, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

To my family for their unconditional love and support.

Abstract

The thesis focuses on the utilization of unstructured data in digital marketing and supply chain management. The first chapter, I explore the effect of expert opinions on consumer experience via the lens of consumer reviews in the restaurant industry, where the expert opinions are conveyed by Michelin stars. We apply two synthetic-control-based methods to estimate the effect of Michelin star changes on the sentiment and content of consumer reviews. We find that decreases in Michelin stars improve consumer review ratings, suggesting that the expectation effect of expert opinions is stronger than the reputation effect. In the second chapter, I move to explore how businesses adapt and respond to expert opinions. To do this, we analyze restaurants historical menus to explore how the restaurants responded to Michelin star awards. We find that one of the reasons why restaurants with decreases in Michelin stars received higher star ratings after the star decrease is that they streamlined their menu structure and thereby improved the service quality. In the final chapter, we conduct a spend analysis of a procurement practice for manufacturers. We propose the three-component classification model to automate spend analysis and replicate the experts know-how. Using the spend data from Cranswick plc, a major food producer in the UK, we demonstrate improved accuracy of our methodology and superior performance compared to benchmark models.

Impact Statement

Text data has emerged as a critical component in the digital era, offering profound insights into consumer behavior, market trends, and business strategies. My research significantly contributes to this burgeoning field of study, highlighting the applications and impacts of text data in digital marketing and supply chain management. This thesis offers a comprehensive approach to understanding and leveraging the nuances of text data, and covers both methodology contributions and practical contributions.

From the methodology perspective, Chapter 2 provides a rigorous and general empirical framework for analyzing consumer response to external shocks with review data by adopting two synthetic-control-based causal inference methods and an augmented LDA model. Chapter 4 presents a comprehensive methodology that employs natural language processing and machine learning to automate spend analysis that successfully replicates the procurement expert's know-how. The categorization model classifies a vast number of suppliers into accurate multi-level hierarchical taxonomy that is both deep and broad.

From the practical perspective, Chapter 4 is the first academic study to formalize the automation of spend analysis. Therefore, it can highlight potential paths and contribute towards the evolution of Industry 4.0. By collaborating with a major food producer in the UK, we demonstrate improved accuracy of our methodology and superior practical performance compared to benchmark models. Our spend analysis decision support tool helps to identify the product categories with the highest sav-

ings potential, and helps recommend specific suppliers to seek savings in an accurate, quick, and cost-effective manner. Simulation of implementation estimates that automation of spend analysis contributes to £16-22 million ($20-28 million) in annual savings for this food manufacturer.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Yiting Deng, for her invaluable guidance, support, and mentorship throughout the duration of my PhD journey. Her expertise and insightful perspectives have been pivotal in shaping my research work. Beyond her role as an advisor, she has been a role model, inspiring me with her dedication and passion for the field. I am immensely grateful for her patience, understanding, and the countless hours she has invested in my growth as a researcher.

I extend my deepest gratitude to my supervisor, Bert De Reyck, for his guidance and unwavering support. Our first meeting, which occurred even before I had started my PhD, marked a decisive turning point; it was his belief in my potential that not only launched the initial phase of this academic journey but also equipped me with the confidence required to tackle its numerous challenges.

I am equally grateful to my co-author Onesun Steve Yoo for his kind and considerate involvement in my research journey. His guidance has been instrumental not only in refining my research but also in shaping my broader career path. I would also like to extend my heartfelt thanks to my co-author, Puneet Manchanda, for his genuine and insightful advice throughout our collaboration.

Special thanks to my friends and fellow Ph.D. students, who have been an integral part of this journey. Whether it was through engaging discussions, shared moments of stress, or celebrating milestones, each one of you has contributed significantly to my personal and professional growth.

Last but not least, words cannot express my gratitude to my family during this difficult journey. The strength and encouragement that flowed from you have been boundless, fueling my resilience and determination to overcome challenges. Your faith in me served as a constant source of motivation, pushing me to persevere even when the path seemed insurmountable. I am particularly thanks to my husband, whose support during numerous sleepless nights provided me with a haven of peace. His sacrifices, often made silently and with great love, have been instrumental in my ability to pursue and fulfill this academic endeavor. To my entire family and my husband, thank you for being my greatest champions and for standing by me every step of the way, providing a foundation of love and support that made all the difference. Your belief in me has been a gift of immeasurable value, and I am forever grateful.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The significance of big data in enhancing decision-making processes and optimizing various business functions, including marketing and supply chain, is well-recognized (Chen et al. 2012a, Davenport 2013, Chae 2015). Despite this, IBM notes that up to 80% of an organization's data is unstructured (George et al. 2014), presenting a vast opportunity for the analysis of such data. The challenge for businesses now lies in harnessing this unstructured data to extract actionable insights, thereby gaining a competitive edge and making more informed decisions. Companies like Amazon, eBay, and Walmart are already utilizing big data text analytics to manage extensive knowledge resources, engage with customers, and improve operational efficiency (Davenport and Patil 2022). This practical application has sparked increasing academic interest in the field. This thesis includes three chapters that offer a comprehensive approach to understanding and leveraging the nuances of text data in digital marketing and supply chain management. To this end, I leverage econometrics models, natural language processing, and machine learning techniques.

In Chapters 2 and 3, I explore how expert opinion affect the responses from demand-side and supply side, respectively. More specifically, Chapter 2 focuses on the demand-side by addressing the effect of expert opinions on consumer experience via the lens of consumer reviews. We answer two research questions: First, how do

changes in Michelin star status affect the sentiment of consumer reviews? Second, how do changes in Michelin star status affect the content of consumer reviews? We construct a unique data set based on the Michelin Guide for Great Britain & Ireland from 2010 to 2020. The data include consumer reviews on TripAdvisor for all restaurants that were awarded Michelin stars during these 11 years, and a large pool of potential control restaurants. We apply two synthetic-control-based methods to estimate the effect of Michelin star changes on the sentiment and content of consumer reviews. We show that decreases in Michelin stars improve consumer review ratings. The analysis of review content further shows that service and "value for money" appear to be the key drivers of the customer experience. When a restaurant loses or receives fewer Michelin stars, consumers become less demanding on service aspects and also focus less on value considerations.

Chapter 3 shifts focus to the responses from supply-side, examining how restaurants have historically adjusted their menus in response to gaining/losing Michelin stars. To do this, we retrieved all available historical menus for each awarded restaurant since the publication of Michelin Guide 2010. We then manually organized the retrieved menus into a structured format. In the empirical studies, we find that one of the reasons why restaurants with decreases in Michelin stars received higher star ratings after the star decrease is that they streamlined their menu structure and thereby improved the service quality. In contrast, restaurants with an increase in Michelin stars tended to focus on menu price rather than the menu structure, which led to complaints about service in consumer reviews.

Chapter 4 focuses on applying natural language processing and machine learning techniques to practical supply chain problems. In collaboration with a UK food manufacturer, I develop a comprehensive methodology to automate spend analysis that successfully replicates the procurement expert's know-how. Automating spend analysis is not straightforward because (i) no true detailed category labels exist for suppliers, (ii) sufficiently large sets of train- ing data do not exist, (iii) hierar-

chical taxonomies vary across manufacturers, and (iv) hierarchical categorization algorithms lose their accuracy beyond two levels. Our comprehensive model overcomes all of these challenges to automate spend analysis and replicate the experts know-how. Using the spend data from Cranswick plc, a major food producer in the UK, we demonstrate improved accuracy of our methodology and superior performance compared to benchmark models. Our spend analysis decision support tool helps to identify the product categories with the highest savings potential, and helps recommend specific suppliers to seek savings in an accurate, quick, and cost-effective manner.

To conclude, this thesis aims to contribute to the methodology and application of utilizing text data in the field of digital marketing and supply chain management. I hope this dissertation to contribute to the expanding body of research that integrates causal inference with machine learning techniques. Additionally, I hope this work inspires other researchers and practitioners who are engaged in applying analytics to address practical decision-making challenges.

Note that I am the first author of all chapters. In all chapters, I performed all analyses and wrote all parts of the chapters myself.

Chapter 2

Expert Opinions on Demand-Side Responses: The Case of Michelin Stars

2.1 Introduction

Customers look to experts and their opinions in their purchase journey as they consider them to be trustworthy sources (Chen and Xie 2005, Johnson et al. 2005, Hilger et al. 2011, Chen et al. 2012b, Friberg and Grönqvist 2012, Ashenfelter and Jones 2013). In the movie industry, 60% of surveyed U.S. consumers stated that movie critic reviews can influence their decision to watch a movie (Statista 2017). In the book industry, the Booker Prize[1] is the leading literary award for the best novel of the year and has been shown to affect book sales (Ginsburgh 2003). In the restaurant industry, the Michelin Guide is one of the best-known and most prestigious expert rating systems, guiding diners in their restaurant choices (Gergaud et al. 2015). Other well-known examples include the American Automobile

[1] https://publishingperspectives.com/2022/03/awards-the-international-booker-prize-names-its-2022-longlist/

Association (AAA)'s Diamond rating in the hotel industry, Robert Parker's Wine Advocate score in the wine industry, and the J.D. Power rating in the automobile industry. Not surprisingly, it has been shown that favorable expert opinions can in general benefit product sales (e.g., Friberg and Grönqvist 2012, Ashenfelter and Jones 2013). However, research on the effect of expert opinions on consumer experience, including post-purchase interactions and consumer evaluations, is relatively sparse. A few notable exceptions include Kovács and Sharkey (2014), Gergaud et al. (2015) and Rossi (2021), who focus specifically on the effect of winning an award on consumer evaluations. Understanding the influence of expert opinions on consumer evaluations is important, because consumer evaluations not only directly reflect their experiences, but also carry tangible behavioral and financial implications, including repeat purchase decisions, revenues and peer recommendations (Mittal et al. 2021, Morgan and Rego 2006).

In this paper, we investigate the effect of both favorable and unfavorable expert opinions on consumer experience. Theoretically, the impact of expert opinions on consumer experience is ambiguous. On the one hand, favorable expert opinions, seen as quality signals, enhance the reputation of the business (e.g., Hilger et al. 2011, Chen et al. 2012b, Ashenfelter and Jones 2013). Consequently, business with their newly gained reputation can potentially witness improved consumer experience driven by consumer conformity, that is, individuals adjust their behaviors or beliefs to align with those of a group or social norm (Asch 1955). We refer to this positive effect of expert opinions as the *reputation effect*. On the other hand, consumer evaluations of their experiences are also based on their expectations in the sense that consumers first have expectations about an experience, then the actual experience, and then they evaluate their experience by comparing it with their expectations. As such, an experience that exceeds/meets/fails to meet their expectations is considered great/good/bad. Past work (e.g., Diehl and Poynor 2010, Fogarty 2012, Gergaud et al. 2015, Sands 2020, Rossi 2021) has shown that endorsements from experts

can influence expectations and as a result influence the final experience (relative to these changed expectations). Therefore, higher expert opinions raise consumer expectations and potentially lead to disappointment as these expectations get harder to meet or exceed, while lower expert opinions can moderate expectations, which then become easier to meet or exceed, leading consumers to be more delighted with their experience. We refer to this effect of expert opinions as the *expectation effect*, noting that it is likely to be a negative effect when expert opinions become more positive. Thus, our objective is to understand the net impact of the reputation and expectation effects of expert opinions on consumer experience through the lens of consumer reviews.

Specifically, we measure the net effect of expert opinions on consumer experience in the restaurant industry, where the expert opinions are conveyed by the Michelin Guide. The Michelin Guide started evaluating restaurants in France in 1900, awarding "stars" to denote quality, and does so worldwide now. We choose this setting for three reasons. First, the restaurant industry has a substantial impact on the economy. According to the National Restaurant Association, the U.S. restaurant industry is forecast to reach $898 billion in sales and provide 14.9 million industry jobs in 2022.[2] Second, the Michelin Guide is updated every year in many countries or regions based on anonymous expert evaluations, with some restaurants added to the list or awarded more stars and others removed from the list or awarded fewer stars. Such annual updates provide us an opportunity to identify the effect of expert opinions summarized (via changes) in awarded Michelin stars. Third, both the reputation effect and the expectation effect have been documented in this context. For example, head chefs describe being awarded a Michelin star as akin to winning an Oscar in Hollywood.[3] In another instance, the Greenhouse restaurant

[2] https://restaurant.org/research-and-media/media/press-releases/association-releases-2022-state-of-the-restaurant-industry-report/

[3] https://www.fcsi.org/foodservice-consultant/worldwide/the-little-red-book/

in London witnessed a 25% increase in diners when it went from one to two Michelin stars.[4] However, Michelin stars can also negatively affect restaurants through heightened consumer expectations. It has been reported that Michelin receives more than 45,000 letters and 7,000 emails from customers every year, and about 25% of these are complaints about unsatisfactory experiences (Johnson et al. 2005). As a chef at a Michelin-starred restaurant stated, "Customers become more demanding, and people expect more of you and criticize things."[5] There have also been cases where the increased pressure and expectations have led chefs to "give back" Michelin stars by revamping their restaurants and food.[6] In fact, this phenomenon - the potential negative impact of Michelin stars - is labelled the "Michelin curse" in the dining industry and food media. We also see evidence in support of this in the consumer reviews we collected.[7]

We construct a unique data set based on the Michelin Guide for Great Britain & Ireland for the years 2010 to 2020. The "Great Britain & Ireland" guide covers two countries - United Kingdom and Ireland - in one book every year. Our dataset consists of 262 restaurants that have been awarded Michelin stars at least once within this time period and 1,257 other "fine-dining" restaurants that never had or received Michelin stars in the same period. We collect consumer reviews for each of these restaurants from TripAdvisor to understand the consumer post-purchase experience

[4] https://www.thestaffcanteen.com/Editorials-and-Advertorials/impact-michelin-stars-business

[5] https://www.bighospitality.co.uk/Article/2017/09/28/Michelin-Guide-chefs-discuss-is-it-still-relevant?

[6] https://www.bbc.com/news/world-us-canada-62854914

[7] For example, in our consumer review data, we find that increasing Michelin stars leads to heightened consumer expectations, e.g., "Bibendum has 2 michelin stars and is very expensive-so our expectations were high..."; "Wouldn't come here again and left feeling annoyed that we had spent £260 on which we felt should have been of a higher standard for 2 michelin stars." Meanwhile, losing Michelin star(s) sometimes leads to improved consumer experiences, e.g.,"This really was the best food I have ever eaten (even compared to a Michelin starred restaurant!)";"The atmosphere is relaxed, friendly, welcoming...a real home from home (unlike some of Edinburgh's other fine dining/Michelin star establishments)."

and evaluations. We focus on TripAdvisor because it is more popular and influential than other platforms (e.g., Google, Facebook, Yelp) for UK consumers.[8] We focus on two kinds of information in these reviews. First, we look at the review sentiment, measured via the five-point scale review rating on overall experience. Second, we analyze the textual content of the consumer reviews in order to get more detailed insights into the rating in the review. For both kinds of review information, we control for "supply" side changes, primarily by restricting our analyses to restaurants that did not change their menu (we collect current and past menus from the restaurant websites) in response to the Michelin star changes.

We apply two synthetic-control-based methods (Abadie et al. 2010, Li 2020) to identify the net effect of Michelin star(s) changes on consumer reviews. In the first method (SCM-DiD), we create a time-varying synthetic control restaurant that best matches the focal awarded restaurant, and then apply the difference-in-differences framework (Hackmann et al. 2015). The second method employs the cohort-based synthetic difference-in-differences (SynthDiD) - see Arkhangelsky et al. (2021) and Berman and Israeli (2022). In terms of the textual analysis of the review data, we extend established Latent Dirichlet allocation (LDA) methods (e.g., Tirunillai and Tellis 2014, Büschken and Allenby 2016, Puranam et al. 2017, Hollenbeck 2018) by allowing for heterogeneous hyper-parameters based on review characteristics and semantic word characteristics.

Setting wise, our work is closest to Gergaud et al. (2015), who show that Michelin stars improve consumers' perceived quality of the awarded restaurant (measured via the Zagat surveys). However, our work differs in three significant ways. First, Gergaud et al. (2015) only consider the *first* publication of the Michelin Guide in a single market (New York City) in 2005. This means that they can only examine the effect of *gaining* Michelin stars. In contrast, we consider the Michelin Guide

[8] https://bdaily.co.uk/articles/2019/06/26/34s-of-uk-consumers-check-online-reviews-tripadvisor-25x-more-influential-than-google

for Great Britain & Ireland during an 11 year period (2010-2020), which allows us to identify all types of changes in Michelin stars and examine the effect of *both gaining and losing* Michelin stars. Second, we rely on consumer reviews, which arise organically rather than through responses to (survey) questions (as that paper does), reducing the potential for bias or distortion associated with survey research, such as sampling bias and non-response bias (e.g., Copas and Li 1997), social desirability response bias (e.g., Krosnick 1999), among others. Consumer reviews also provide deeper insights as we are able to analyze both the rating and the associated text. Third, Gergaud et al. (2015) use difference-in-differences and propensity score matching methods. We are able to leverage state-of-the-art methods in causal inference - two synthetic-control-based methods - that provide better identification, especially in terms of controlling for time-varying confounders (cf. Xu 2017).

Our results on review sentiment show that decreases in Michelin star(s) improve the consumer review ratings, suggesting that the expectation effect of expert opinions is stronger than the reputation effect. In contrast, an increase in Michelin star(s) has no impact on the consumer review ratings. Turning to the analysis of the review content data, we find that when a restaurant loses or receives fewer Michelin stars, consumers become less demanding on service aspects and also focus less on "value for money" considerations. In addition, consumers also appear less concerned about food in their reviews. These results are consistent across both synthetic-control-based methods. We demonstrate that these results are unlikely driven by supply-side responses to Michelin awards or demand-side responses unrelated to the expectation and reputation effects, such as the changes in the mix of consumers visiting the restaurant. We also show the robustness of these results via an analysis that uses observable restaurant characteristics to select the control group, analyses with an alternative dependent variable and an alternative time window, and an additional falsification test.

Our findings go some way in terms of shedding light on the "Michelin curse."

The Michelin Guide has five publicly acknowledged assessment criteria: quality of the products, mastery of flavor and cooking techniques, the personality of the chef in the cuisine, value for money, and consistency between visits.[9] In order to gain and/or keep a Michelin star, restaurants need to perform to satisfy these criteria. Many chefs struggle with these, especially consistency as that dampens creativity and lowers innovation. In fact, according to Hayward (2021), Michelin awards "damage" restaurants, causing them to narrow their creativity to obtain stars and to stop innovating in order to keep the stars. Overall, our paper suggests that losing Michelin stars is not necessarily bad news for restaurants, especially vis-a-vis the consumer experience. Conversely, winning Michelin stars does not seem to improve the customer experience in any material way.

To summarize, this paper makes the following contributions. First, we conduct a rigorous analysis on the (net) effect of expert opinions on consumer experiences. Our findings show that, in our setting, a lower rating from experts can lead to a better consumer experience, i.e., the expectation effect dominates the reputation effect. To the best of our knowledge, this is the first instance of the documentation of this outcome. Second, by analyzing consumer review text data, we identify key drivers of the customer experience, further enriching managerial insights on the value of receiving favorable or unfavorable expert opinions. Third, by adopting two synthetic-control-based causal inference methods and an augmented LDA model, we provide a rigorous and general empirical framework for analyzing consumer response to external shocks with review data. Finally, we provide a data-based explanation for the "Michelin curse," offering implications for chefs, restaurant managers, the Michelin Guide, and other businesses that provide experience goods.

The rest of the chapter is organized as follows. We describe the data and present descriptive statistics in Section 2.2, followed by the empirical strategy in Section 2.3 and empirical results in Section 2.4. We test multiple alternative explanations

[9] `https://guide.michelin.com/en/article/news-and-views/how-to-get-michelin-stars`

in Section 2.5 and conduct robustness checks in Section 2.6. Finally, we conclude in Section 2.7 with a discussion of the managerial implications, limitations and potential future extensions.

2.2 Data

2.2.1 The Michelin Guide and Awarded Restaurants

The Michelin Guide evaluates restaurants via the use of a group of anonymous inspectors that operate worldwide. Inspectors are anonymous when visiting the potential restaurants in order to guarantee that restaurants treat them as regular consumers.[10] Every decision relating to Michelin stars is decided by multiple inspectors from different global regions who take turns to visit a restaurant in order to ensure that the final outcome is based on a consensus (among inspectors). In other words, no single inspector can assign or remove Michelin stars for a restaurant.

We construct a comprehensive data set based on the Michelin Guide for Great Britain & Ireland from the year 2010 to 2020. We denote restaurants which received Michelin stars at least once within this time frame as *awarded restaurants*. For each awarded restaurant, we extract the restaurant's characteristics (e.g., official website URL, address, postcode, price level and cuisine type) from the restaurant's TripAdvisor page, and use a postcode checker to identify whether the restaurant is in an urban or a rural area.[11] In total, our data cover 262 awarded restaurants

[10] See, for example, `https://www.forbes.com/sites/karlaalindahao/2019/10/23/the-secret-life-of-an-anonymous-michelin-restaurant-inspector-2019/?sh=230efd5135c9`

[11] The rural/urban classification is based on offices of national statistics in the UK (i.e., England, Wales, Scotland, Northern Ireland) and Ireland. In England and Wales, the rural/urban classification was developed by the Office for National Statistics. In Scotland, the rural/urban classification was developed by the Scottish governments Geographic Information Science & Analysis Team. In Northern Ireland, the rural/urban classification was published by the Northern Ireland Statistics and Research Agency (NISRA). In Ireland, an interactive map, "area type classification," was developed by the Central Statistics Office.

that received Michelin stars at least once.[12] Among these awarded restaurants, 91 (34.7%) are located in London, 235 (89.7%) are associated with the highest price level (as labeled by TripAdvisor), 234 (89.3%) specialize in European cuisines (e.g., British, European and French, etc.), and 170 (64.9%) are located in urban areas.

As our goal is to analyze the effect of expert opinions (conveyed via Michelin stars) on consumer reviews, it is crucial to observe Michelin star changes (either an increase or a decrease in the number of Michelin stars, or an addition to or a deletion from the Michelin Guide). We define the *guidebook year* as the period between the publication dates of two consecutive guides. For example, the 2019 Michelin Guide was published on October 1, 2018 and the 2020 Michelin Guide was published on October 7, 2019, so the period between these two dates corresponds to guidebook year 2019. During these 11 guidebook years, 207 (79.0%) awarded restaurants experienced Michelin star changes at least once, and the remaining 55 (21.0%) restaurants kept the same Michelin stars throughout.

Table 2.1 lists the Michelin star awards and Michelin star changes by guidebook year. Every guidebook includes more than one hundred awarded restaurants, most of which are one-star restaurants. Michelin star increases can be new additions to the Michelin list (e.g., from no-star to one-star) or gaining more stars (e.g., from one-star to three-star), and Michelin star decreases can be removals from the Michelin list (e.g., from one-star to no-star) or losing stars but remaining on the list (e.g., from three-star to one-star). In total, there are 269 star changes, with 174 star increases and 95 decreases. In this paper, we do not separate the cases within Michelin star increases and Michelin star decreases, because we only observe 15 (out of 269) instances where an awarded restaurant gained more stars, and 6 (out of 269) instances where an awarded restaurant lost stars but remained on the list. Figure 2.1 visualizes the Michelin star transitions by year.

[12] We exclude twelve (out of 278) restaurants that did not have a TripAdvisor page, and four (out of 278) that did not have an official website.

Figure 2.1: Michelin Stars Transition by Year

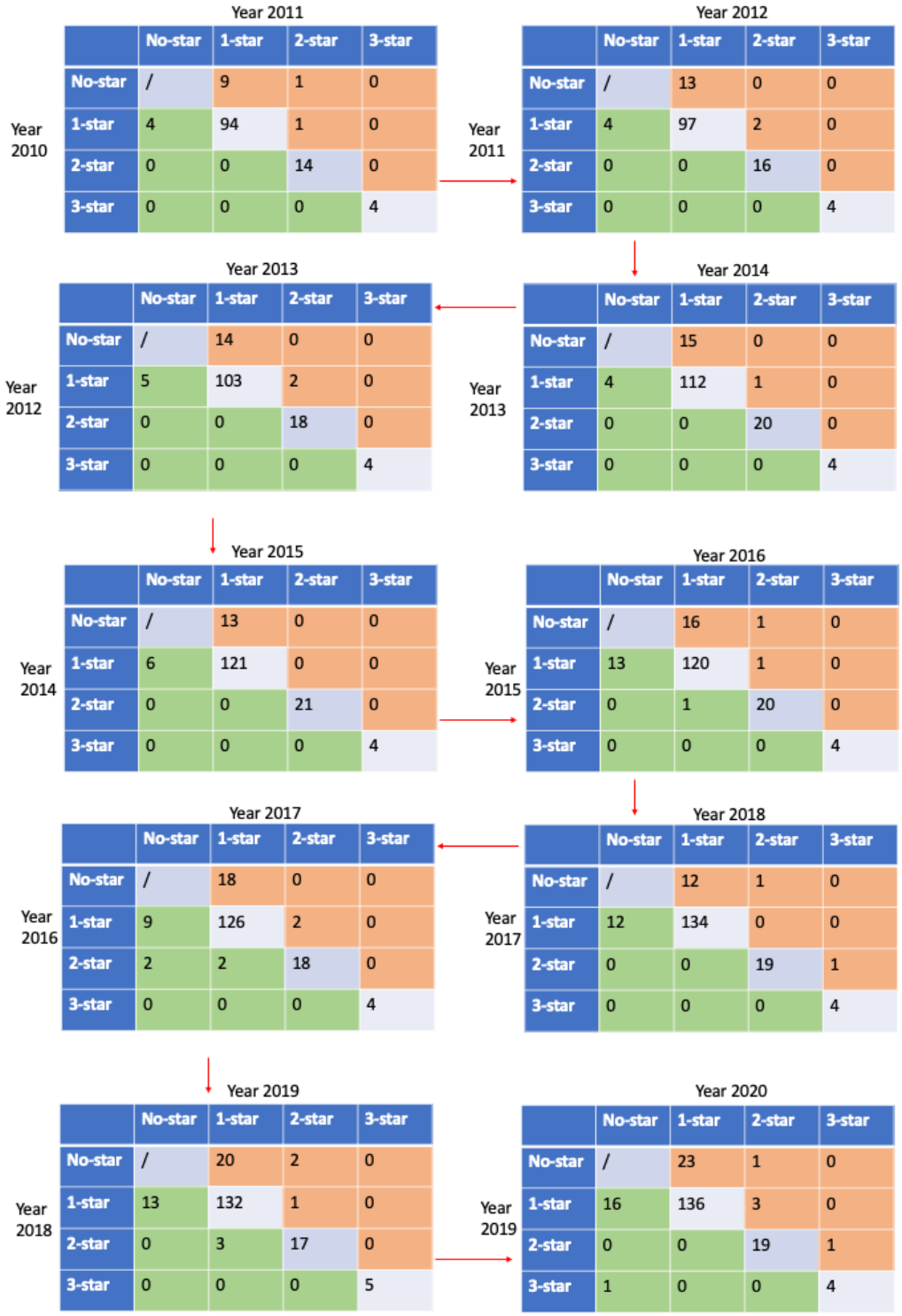


Table 2.1: Summary of Michelin Stars (2010 to 2020)

Guidebook year	# Michelin restaurants				# Michelin star changes				
	total	one-star	two-star	three-star	total	increase– additions to the guide	increase– awarded but gaining more stars	decrease– removals from the guide	decrease– losing stars but remaining on the guide
2010	117	99	14	4	–	–	–	–	–
2011	123	103	16	4	15	10	1	4	0
2012	132	110	18	4	19	13	2	4	0
2013	141	117	20	4	21	14	2	5	0
2014	152	127	21	4	20	15	1	4	0
2015	159	134	21	4	19	13	0	6	0
2016	163	137	22	4	32	17	1	13	1
2017	170	146	20	4	33	18	2	11	2
2018	171	146	20	5	26	13	1	12	0
2019	180	155	20	5	39	22	1	13	3
2020	187	159	23	5	45	24	4	17	0

Note: We do not consider Michelin star changes in guidebook year 2010 because guidebook year 2010 provides the initial star levels for the period under investigation.

2.2.2 Pool of Control Restaurants

For identification purposes, we further construct a large pool of *control restaurants*, which never received Michelin stars during the data period, are located in the cities with at least one awarded restaurant and are categorized as "fine-dining" on TripAdvisor.[13]

Specifically, we take the following steps to construct the pool of control restaurants. First, we check the city information for each awarded restaurant on Google, and then collect the city's TripAdvisor restaurant page. The 262 awarded restaurants are located in 73 cities in Great Britain and Ireland. Second, from the city's TripAdvisor restaurant page, we scrape the URL links of all "fine-dining" restaurants that are not in the list of the 262 awarded restaurants. We are able to collect the TripAdvisor URLs of 1,803 "fine-dining" restaurants in total, which includes 262 awarded restaurants and 1,541 "fine-dining" restaurants that never received Miche-

[13]On TripAdvisor, restaurants are assigned one of the three labels: "Cheap-eats," "Mid-range," and "Fine-dining." Fine-dining restaurants are typically associated with the price-level symbol "££££", though in rare instances, some are marked with the symbol "£££" or "££" Specifically, 27 (out of 262) awarded restaurants and 60 (out of 1,257) control restaurants are fine-dining with "££" or "£££" price labels.

lin stars during the data period (i.e., control restaurants). Third, similar to data collection for awarded restaurants, we collect information on each of these 1,541 control restaurants, including their official website URL, characteristics (e.g., address, postcode, price level and cuisine type), and rural/urban classifications. 284 (out of 1,541) control restaurants did not receive consumer reviews during the period of our study, and are dropped from the control pool. Thus, we have a pool of 1,257 control restaurants, of which 1,197 (95%) are associated with the highest price level (as denoted by TripAdvisor).

2.2.3 Consumer Review Data

We scrape TripAdvisor consumer reviews for each of the 262 awarded restaurants and the 1,257 control restaurants. As discussed earlier, TripAdvisor is chosen because it is more popular and influential than other platforms (e.g., Google, Facebook, Yelp) for UK consumers. The consumer reviews include the review text and an overall evaluation of the dining experience on a five-point scale, with a higher rating indicating a better experience. Our sample includes 889,660 consumer reviews.

Table 2.2 reports key statistics on the review data by Michelin star level. Note that a single awarded restaurant can appear with different Michelin star levels in different years. Overall, holders of higher Michelin stars have more consumer reviews on TripAdvisor. This is likely due to the reputation effect of Michelin stars: consumers are more likely to visit, review, and indicate their satisfaction (or not) with an awarded restaurant. While the consumer review ratings for the awarded restaurants are somewhat higher than those for the control restaurants, the differences are not statistically significant.

Table 2.2: Summary Statistics of the Review Data (by Michelin star level)

	Awarded Restaurants				Control
	No-star	One-star	Two-star	Three-star	Restaurants
Number of restaurants	252	241	31	6	1,257
Number of reviews	46,044	146,683	35,445	7,521	653,967
Avg number of reviews per restaurant	183	609	1143	1,254	520
Mean of restaurant-level average review rating (s.d.)	4.50 (0.40)	4.47 (0.28)	4.58 (0.22)	4.63 (0.25)	4.25 (0.60)

Note: "No-Star" refers to awarded restaurants in guidebook years when they did not receive a Michelin star. "Control Restaurant" refers to restaurants that never received Michelin stars in the data period.

2.2.4 Use of Menus as Supply-Side Controls

Changes in Michelin star status could result in restaurants adjusting various aspects such as food, decor, service etc. As mentioned earlier, we control for these via our sample construction. First, we retrieve all available historical menus for each awarded restaurant and control restaurant since the publication of the Michelin Guide 2010, using the Wayback Machine (`https://archive.org/web/`) to access archived versions of the restaurants' official websites. Then, for each restaurant, we check menus on each date that the website has been archived,[14] and determine whether there have been any changes compared to the last archived menu. Over the 11-year period (2010 - 2020), we find that the number of menu changes is quite modest, averaging 15.8 changes for an awarded restaurant and 5.1 for a control restaurant. In order to control for menu changes, we restrict our data to include only those awarded restaurants and control restaurants *without* menu changes in the 180-day period around the Michelin Guide release (90 days before and 90 days after the publication date). As a result, we exclude 17 (out of 269) star change observations in the awarded group and 110 (out of 1,257) restaurants in the control

[14]Note that the Wayback Machine does not archive all websites on a daily basis.

group.[15]

Second, the restriction of the time window to just 90 days post the Michelin Guide release makes it unlikely that restaurants can successfully make major (non-menu) changes, e.g., decor and/or re-training the staff to deliver a different service level. In addition, we carry out a detailed analysis of the occurrence of service related topics in consumer reviews (for each restaurant type - star increase, star decrease, no star change, control) in each twelve-month period after the release of the Michelin guide. We find that the attention paid to service (in the reviews) stays stable. Section 2.5.1 provides the relevant details supporting our conjecture that supply-side changes do not drive our results.

2.2.5 Final Sample in Main Analyses

After making the above selections, the final sample we use in the following empirical analysis includes 252 star changes (denoted as *treated unit*) and 1,147 control restaurants. Table 2.3 shows the number of awarded restaurants gaining Michelin stars, the number of awarded restaurants losing Michelin stars, and the number of control restaurants in the pool, by guidebook year. Note that not all restaurants have received consumer reviews every year, so the number of control restaurants varies by year and generally increases over time because of consumer review accumulation.

2.2.6 Additional Reviewer-level Data

To analyze whether changes in Michelin stars change the mix of consumers who visit the restaurant (e.g., Bondi et al. 2023), we further collect comprehensive data about the reviewers, as outlined below.

First, to understand if the restaurant attracts different types of consumers after

[15] In robustness checks not reported in the paper, our main findings remain consistent without controlling for menu changes at the awarded and control restaurants. Results are available upon request from the authors.

Table 2.3: Summary of the Number of Restaurants in Empirical Analyses (by Guidebook Year)

Guidebook year	# Awarded restaurants with Michelin star increases	# Awarded restaurants with Michelin star decreases	# Control restaurants
2011	11	4	429
2012	14	3	549
2013	16	5	599
2014	13	4	648
2015	13	5	716
2016	17	13	770
2017	20	12	830
2018	14	11	910
2019	23	13	1,001
2020	28	13	1,065
Total	169	83	7,517

Note: As a control restaurant can be included in the control pool for multiple guidebook years, the sum of the control restaurants units exceeds the total number of 1,147.

the Michelin star change, we collect the TripAdvisor profile pages of reviewers who have reviewed an awarded restaurant within the 90-day guidebook window. The TripAdvisor profile page contains reviewer-level information, such as their location of registration, registration time, and all of the reviews they have posted (not limited to those for the awarded restaurants). We collected TripAdvisor profile pages for 52,210 unique reviewers, who have written 1,617,923 reviews from 2010 to 2020.

Second, we collect restaurant information associated with these 1,617,923 reviews. These reviews are associated with 327,852 unique restaurants. For each of these 327,852 restaurants, we access its TripAdvisor page to collect restaurant characteristics and all consumer reviews. We were able to locate TripAdvisor pages for 279,359 (out of 327,852) restaurants. These restaurants have been reviewed by 45,274 (out of 52,210) reviewers in the data, and have received a total number of over 79 million reviews. These review data will enable us to assess whether changes in Michelin stars led consumers to visit a different type of restaurant.

2.3 Empirical Strategy

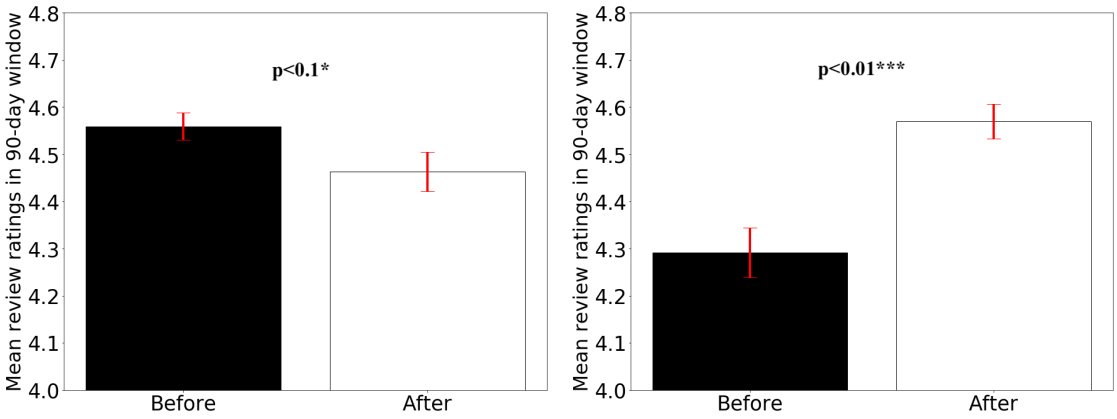
This section proceeds as follows. First, to provide model-free evidence, Section 2.3.1 shows the mean review ratings for treated units in the 90-day windows before and after Michelin star changes, respectively for those gaining stars and for those losing stars. Next, we describe two variants of the synthetic control method (SCM) for estimating the causal impact of Michelin star changes.

Section 2.3.2 describes the first method SCM-DiD (Hackmann et al. 2015) and Section 2.3.3 describes the second method SynthDiD (Arkhangelsky et al. 2021, Berman and Israeli 2022). Both SCM-DiD and SynthDiD have been shown to provide clean identification and aid in causal inference. Each method has its own advantages. SCM-DiD creates a time-varying synthetic control restaurant to best match each treated unit and then applies the difference-in-differences framework, allowing controls for fixed effects at the restaurant level and the time level. SynthDiD separates treated units into guidebook-specific treated cohorts and then employs the cohort-based synthetic difference-in-differences model, which relaxes the strong parallel-trend assumptions for all units and all time periods. We apply both methods to ensure the robustness of the results.

2.3.1 Model-Free Evidence

Figure 2.2 shows the mean review ratings received by the awarded restaurants in the 90-day windows before and after the Michelin star changes. Clearly, the restaurants with Michelin star increases (left panel) received lower consumer review ratings after the Michelin star changes, and the restaurants with Michelin star decreases (right panel) received higher consumer review ratings after the Michelin star changes. The initial model-free evidence suggests a relationship between the Michelin star changes and the consumer review ratings, which is in line with the expectation effect.

Figure 2.2: Mean Review Ratings in a 90-day Window Before/After Guidebook Release for Michelin Star Increases (left) and Michelin Star Decreases (right)



Note: Error bars represent standard deviations.

This pattern obviously does not control for potential confounding factors. Table 2.4 summarizes identification challenges, possible confounding factors and alternative explanations for the observed pattern(s), along with our approach to dealing with these.

2.3.2 Synthetic Control Method and Difference-in-Differences Framework (SCM-DiD)

After the release of a new guidebook, an awarded restaurant is either treated (i.e., with Michelin star changes) or untreated (i.e., without Michelin star changes). In order to predict the potential outcomes of a treated unit "as if" there were no Michelin star changes, we employ the synthetic control method (SCM, Abadie et al. 2010, 2015) to create a best-matching control restaurant before its Michelin star change. The synthetic control method allows us to capture any possible trends that might affect identification of the effect of the Michelin star change.

For each treated unit (i.e., an awarded restaurant with a Michelin star change in a specific guidebook year), we create a donor pool which consists of all available control restaurants offering the same type of cuisine. Then, for the focal awarded restaurant

and each control restaurant in the donor pool, we construct a "restaurant-guidebook year" panel of consumer reviews with the following variables: yearly average review ratings, yearly variance of review ratings, and yearly cumulative number of reviews.

Based on the "restaurant-guidebook year" panel data, we construct a synthetic control restaurant for each treated unit as a weighted combination of the donor restaurants, with weights chosen so that the resulting synthetic control restaurant best-approximates the relevant characteristics of the treated unit in the pre-treatment period. On average, a synthetic control restaurant is constructed from a pool of 404 control restaurants. The outcome variable is the yearly average review rating. The predictors include the yearly variance of review ratings, yearly cumulative number of reviews, and price level. In addition, we follow Abadie et al. (2011) to include as a special predictor, the average review rating in the 90-day pre-treatment period, to ensure that the synthetic restaurant is similar to the treated unit right before the treatment. For every restaurant, the outcome variable and predictors are calculated with an average number of 601 reviews. Therefore, we constructed 223 synthetic control restaurants corresponding to 223 (out of 252) treated units. The remaining 29 treated units do not have enough reviews on at least one side of the treatment time and therefore are dropped in the SCM procedure.

Next, we undertake an event study approach and focus our analysis on a window of 90 days before (pre-guidebook window) and 90 days after (post-guidebook window) the release of the new guidebook. The SCM procedure described above results in 223 pairs of treated and synthetic control restaurants. For each pair of treated unit and its synthetic control, we aggregate the reviews and retain observations in the pre- and post- windows, so that there are four observations on each pair: treated-pre, treated-post, control-pre, and control-post.

Finally, we estimate the effects of Michelin star changes on consumer reviews in a stacked difference-in-differences framework (Hackmann et al. 2015). When examining the effect on the review sentiment, we use the mean consumer review rating as

the dependent variable. When analyzing the effect on the review content, we first extract topics from text reviews and then use the mean probability of each topic as the dependent variable. The stacked difference-in-differences model is specified as follows:

$$\begin{aligned} Y_{it} = \quad & \beta_1 After_{it} + \beta_2 After_{it} \times Increase_{it} + \beta_3 After_{it} \times Decrease_{it} \\ & + \beta_4 OneStar_{it} + \beta_5 TwoStar_{it} + \beta_6 ThreeStar_{it} \\ & + \beta_7 X_{it} + \beta_8 Z_{it} + \alpha_{p(i)w(t)} + \gamma_i + \varepsilon_{it} \end{aligned} \tag{2.3.1}$$

where i denotes restaurant, t denotes guidebook year, and $w(t)$ denotes the guidebook window defined as a window of 90 days before and 90 days after the release of guidebook for year t ($t \in \{2011, ..., 2020\}$). Therefore, the guidebook window $w(t)$ includes observations in guidebook year $t-1$ and observations in guidebook t.

The dependent variable, Y_{it}, is the outcome of interest (e.g., mean review rating in sentiment analysis, mean topic probabilities in content analysis) for restaurant i in the part of the guidebook window belonging to guidebook year t. $After_{it}$ is an indicator variable which takes the value of 1 if the observation is in the post-guidebook window, and takes the value of 0 otherwise. We include dummy variables - $Increase_{it}$ and $Decrease_{it}$ - to denote two treatment groups, indicating the changes in Michelin star (i.e., increase, decrease, or unchanged). Specifically, $Increase_{it}$ ($Decrease_{it}$) takes the value of 1 if restaurant i gained (lost) stars in guidebook year t compared with guidebook year *t–1*. The interaction term between $After_{it}$ and $Increase_{it}$ ($Decrease_{it}$) therefore measures the treatment effect above and beyond the general trend. Corresponding to the three-star rating system in the Michelin Guide, we add three indicator variables, $OneStar_{it}$, $TwoStar_{it}$, and $ThreeStar_{it}$, to control for the current Michelin star level of restaurant i in guidebook year t. X_{it} is a vector of cumulative review characteristics for restaurant i in the window

belonging to guidebook year t, constructed based on all available reviews prior to the window. These characteristics are: the logarithm of the total number of reviews, the cumulative average review rating, and the variance of previous ratings. Z_{it} is a measure of average demand of restaurant i in the 90-day window, proxied by the normalized search intensity collected from Google Trends. We include pair–window fixed effect $\alpha_{p(i)w(t)}$ to control for unobservable factors affecting the restaurant pair $p(i)$ during the window $w(t)$. Restaurant fixed effect γ_i controls for unobservable time-invariant restaurant characteristics such as the restaurant's general decoration style, and ε_{it} is an idiosyncratic error term.

Table 2.4: Identification Challenges, Alternative Explanations and Proposed Solutions

Type			Solutions and Empirical Models
Identification Challenges		General trend	*SCM-DiD* (Section 2.3.2) and *SynthDiD* (Section 2.3.3). Robustness check using placebo guidebook publication date (Section 2.6.4).
		Different panel lengths across restaurants before treatment	*SynthDiD* Use 18-month review data for each treated and control unit (Section 2.3.3).
		Control restaurants selected based on SCM may not be fully comparable with treated restaurants	*SCM-DiD* Robustness check using manually selected control restaurants based on location, price, and cuisine type (Section 2.6.1).
Alternative Explanations	Supply-side changes	Restaurants change menus	*SCM-DiD* and *SynthDiD* Focus on restaurants without menu changes in the 180-day period around the Michelin Guide publication date (Section 2.2.5).
		Restaurants make changes on serving size (even with same food)	*SCM-DiD* and *SynthDiD* Robustness check using subset of restaurants evidencing consistency (Section 2.5.1).
		Restaurants make major non-food changes (e.g., decor, service)	*SCM-DiD* Focus on the short time window around the Michelin Guide publication date (Section 2.3.2). Robustness check using an alternative time window (Section 2.6.3). *Other Analysis* Focus on "service-related topic metrics, and analyze probabilities of relevant topics over the twelve-month period between guidebook releases (Section 2.5.1).
	Demand-side changes	Restaurant demand changes	*SCM-DiD* and *SynthDiD* Use log-transformed normalized Google search intensity as the dependent variable. Use Farronato and Zervas (2022)'s OpenTable reservation as the dependent variable (Section 2.5.2).
		Consumers show sympathy towards restaurants losing stars	*SCM-DiD* and *SynthDiD* Use review volume as the dependent variable (Section 2.5.2). Replication with restaurants serving British cuisine (Section 2.5.2).
		Selection of different consumers visiting the restaurant	Analyze whether a restaurant attracts different types of consumers after the Michelin star change, and whether a Michelin star change led consumers to visit a different type of restaurant. (Section 2.5.2).
		Michelin star changes may change the proportion of extreme reviews	*SCM-DiD* and *SynthDiD* Use the percentage of 5-star reviews to measure restaurant-level sentiment (Section 2.6.2).

2.3.3 Synthetic Difference-in-Differences (SynthDiD)

The SCM-DiD model presented above includes 223 synthetic control restaurants, one for each treated unit. As our data span 11 guidebook years, these synthetic control restaurants may have different panel lengths before treatment, depending on the guidebook year of Michelin star changes. Different pre-treatment panel lengths in SCM may bias the estimates, thus we address this potential issue with the synthetic difference-in-differences (SynthDiD) approach (Arkhangelsky et al. 2021). SynthDiD allows both unit and time weights, where the unit weights are selected in a similar way as SCM, and time weights are added so that within a unit, the weighted average outcomes across pre-treatment periods approximate those in the post-treatment period.

The SynthDiD is designed for a balanced panel where the treated units have the same treatment time. In our setting, treatment time varies by restaurant. Therefore, we follow Berman and Israeli (2022) to adapt the SynthDiD method to the staggered treatment time by separating treated units into guidebook-specific treated cohorts, estimating the treatment effect for each cohort separately, and then aggregating them into an overall average treatment effect. We do this in four steps. First, for each guidebook year t, we create three cohorts: treated cohort $r_t^{increase}$ consisting of treated units with an increase in Michelin stars; treated cohort $r_t^{decrease}$ consisting of treated unites with a decrease in Michelin stars; and control cohort $r_t^{control}$ consisting of control restaurants. We denote the number of restaurants in the three cohorts respectively by $N_t^{increase}$, $N_t^{decrease}$, and $N_t^{control}$.

Second, for each restaurant in the guidebook-specific treated or control cohort, we extract review data in the period of one year before treatment to six months after treatment. We then divide the 18-month data into nine consecutive two-month blocks, and calculate the restaurant-level mean outcome (e.g., review rating in sentiment analysis, topic probabilities in content analysis) in each two-month

block. A restaurant is excluded from the cohort if it does not have the full nine blocks of data, or if it is an awarded restaurant but has more than one change of Michelin stars within this 18-month period (i.e., changed Michelin stars in two consecutive years). As a result, we retain 148 (out of 252) treated units, including 95 treated units for gaining Michelin stars (*Increase*), and 53 treated units for losing Michelin stars (*Decrease*). Correspondingly, there are 4,334 control units. Table 2.5 summarizes the treated and control cohorts in the data constructed above.

Third, for each guidebook year t, we estimate the cohort-level treatment effect of gaining Michelin stars, $ATT_t^{increase}$, using treated cohort $r_t^{increase}$ and control cohort $r_t^{control}$. Similarly, we estimate the cohort-level treatment effect of losing Michelin stars, $ATT_t^{decrease}$, using treated cohort $r_t^{decrease}$ and control cohort $r_t^{control}$. Standard errors are estimated using bootstrapping (Algorithm 2 of Arkhangelsky et al. (2021)), or the placebo method (Algorithm 4 of Arkhangelsky et al. (2021)) if a cohort includes only one treated restaurant.

Table 2.5: Summary of Treated and Control Cohorts in SynthDiD

	Treated Cohorts		Control Cohorts
	Increase	Decrease	
Total number of units	95	53	4,334
Avg. number of units in a guidebook-specific cohort	10.6	5.9	481.6
Avg. number of reviews per unit (within 18 months)	175.1	143.2	194.0

Note: As a control restaurant can be included in the control pool for multiple guidebook years, the sum of the control units exceeds the total number of control restaurants 1,147.

Lastly, we aggregate the cohort-level treatment effects to the overall treatment effect (ATT) by taking the weighted average as follows:

$$ATT^{increase} = \frac{\sum_t N_t^{increase} \cdot ATT_t^{increase}}{\sum_t N_t^{increase}} \tag{2.3.2}$$

$$ATT^{decrease} = \frac{\sum_t N_t^{decrease} \cdot ATT_t^{decrease}}{\sum_t N_t^{decrease}} \tag{2.3.3}$$

Standard errors for the overall ATT are computed as a weighted average of the cohort-level standard errors.

2.4 Results

This section reports the estimation results from the SCM-DiD and SynthDiD analyses. Section 2.4.1 reports the effects of Michelin star changes on sentiment of consumer reviews. Section 2.4.2 reports the results on content of consumer reviews, where we first extract topics of consumer reviews using the Latent Dirichlet Allocation (LDA) model (Section 2.4.2), and then estimate the effect on topic probabilities (Section 2.4.2).

2.4.1 Effects of Michelin Stars on Sentiment of Consumer Reviews

Table 2.6 presents the results of the SCM-DiD model (Section 2.3.2), using the mean consumer review rating as the dependent variable in Equation (2.3.1). Column (1) controls only for Michelin star levels and fixed effects, and Column (2) adds the full set of controls. The estimated coefficient for $After$ is significantly negative, suggesting a declining trend in online ratings, which is consistent with prior literature (e.g., Moe and Trusov 2011, Li and Hitt 2008). The estimated coefficient for $After \times Increase$ is insignificant, suggesting that gaining Michelin stars does not lead to changes in consumer review ratings. However, the estimated coefficient for $After \times Decrease$ is significantly positive and the magnitude is larger than that of $After$, suggesting an increase in the consumer review ratings for restaurants that lost Michelin stars. This is likely driven by the expectation effect of Michelin stars: consumers lower their expectations for restaurants with Michelin star decreases, and as a result, tend to be more satisfied with the dining experience.

Table 2.6: Effects of Michelin Star Changes on Sentiment of Consumer Reviews by SCM-DiD

	Synthetic control + Difference-in-Differences	
	(1)	(2)
After	–0.069***	–0.045**
	(0.014)	(0.020)
After × Increase	0.040	0.049
	(0.052)	(0.056)
After × Decrease	0.318***	0.281***
	(0.058)	(0.064)
One Star	–0.060	–0.072
	(0.048)	(0.057)
Two Star	0.002	–0.024
	(0.081)	(0.102)
Three Star	0.998***	0.955***
	(0.196)	(0.196)
ln(number of reviews+1)		–0.056**
		(0.028)
Cumulative average rating		0.654*
		(0.365)
Cumulative rating variance		0.108
		(0.353)
ln(normalized search volume+1)		–0.017
		(0.103)
Pair–window FE	Yes	Yes
Restaurant FE	Yes	Yes
Observations	892	892
Number of pairs	223	223
R^2	0.851	0.856

Note: Robust standard errors clustered at pair level are in parentheses, and they are Diff-in-Diff regression-based clustered standard errors. In Online Appendix 2.8.1, we report bootstrapped standard errors following Arkhangelsky et al. (2021) and Adalja et al. (2023). The results remain consistent. *** p<0.01,** p<0.05,* p<0.1.

Table 2.7 presents the results of the SynthDiD model (Section 2.3.3).[16] The overall ATT for Michelin star increases remains insignificant, and the overall ATT for Michelin star decreases is significant with a value of 0.311. Based on the 18-month data around the treatment time, we construct Figures 2.3 and 2.4, which

[16]The detailed ATT estimates by cohort can be found in Appendix 2.8.2.

show the dynamic treatment effect estimates (with their 95% confidence intervals) for gaining and losing Michelin stars respectively.[17] Time 1 represents the first two-month block after the Michelin star changes, and other times represent two-month blocks relative to the Michelin star change. In the pre-treatment periods (i.e., Time –5 to Time 0), the estimated ATT values are approximately zero in both figures, confirming the parallel pre-trends. In the post-treatment periods (i.e., Time 1 to Time 3), Figure 2.3 shows the confidence intervals on the ATT for gaining Michelin stars contain zero in all periods, suggesting that the consumer review ratings do not change after the restaurant gained Michelin stars. In Figure 2.4, it is evident that the ATTs for losing Michelin stars are positive, indicating an increase in consumer review ratings for restaurants that lost Michelin stars. Both plots are consistent with the overall ATTs reported in Table 2.7.

Table 2.7: Effects of Michelin Star Changes on Sentiment of Consumer Reviews by SynthDiD

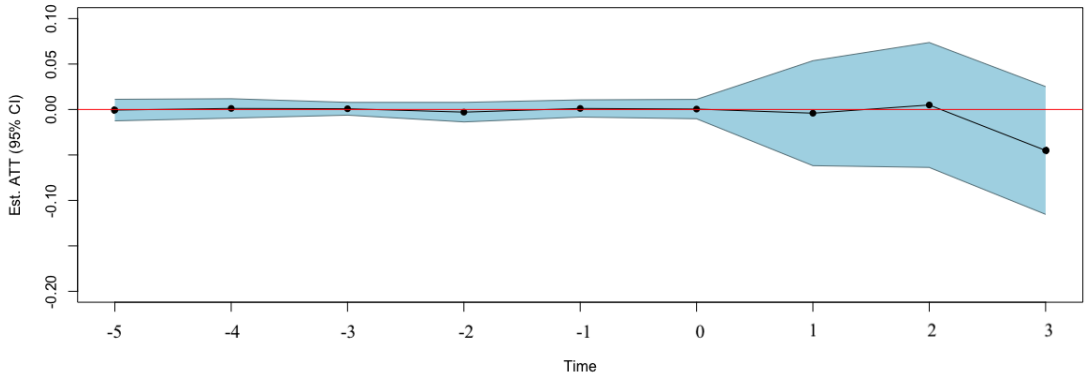
	Aggregated Synthetic Difference-in-Differences	
	(1) Increase	(2) Decrease
Overall ATT	–0.015	0.311**
	(0.064)	(0.138)
Total number of treatment units	95	53
Total number of control units	4,334	4,334

Note: Aggregated standard errors are in parentheses. ** p<0.05.

Together, our results show that when a restaurant loses stars, the positive expectation effect outweighs the negative reputation effect, leading to higher consumer ratings. In contrast, when a restaurant gains stars, it is possible that the positive reputation effect negates the potential negative effects of higher expectations, leading to an overall null effect. However, we are unable to separate the expectation

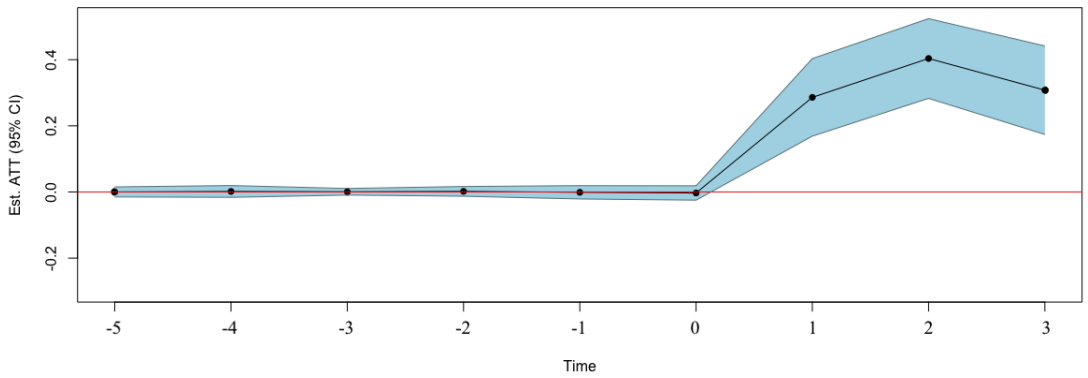
[17] See Appendix A.1 in Berman and Israeli (2022) for details on the methodology used to create the plot.

Figure 2.3: SynthDiD Treatment Effects for Gaining Michelin Stars



Note: Time –5 to Time 0 correspond to the one-year pre-treatment period (i.e., six consecutive two-month blocks), and Time 1 to Time 3 are post-treatment periods (i.e., three consecutive two-month blocks). Time 1 denotes the first two-month block after the Michelin star increase.

Figure 2.4: SynthDiD Treatment Effects for Losing Michelin Stars



Note: Time –5 to Time 0 correspond to the one-year pre-treatment period (i.e., six consecutive two-month blocks), and Time 1 to Time 3 are post-treatment periods (i.e., three consecutive two-month blocks). Time 1 denotes the first two-month block after the Michelin star decrease.

effect from the reputation effect given the observational nature of our data.

We further check whether the results vary with the restaurant location, price level or cuisine type, but do not find any such differences.

2.4.2 Content of Consumer Reviews

Having demonstrated the effect of Michelin star changes on the consumer review ratings, we next delve into the content or the reviews to understand the mechanisms behind the effects. We apply an LDA model to extract topics from textual reviews (Section 2.4.2), and estimate the effects of Michelin star changes on the identified

topics using SCM-DiD and SynthDiD (Section 2.4.2).

LDA Model

To facilitate textual data analysis, we preprocess all reviews by splitting the text into its component words, eliminating punctuations, lemmatizing words into dictionary form, transforming plurals to singular, removing stop words (a, an, the, etc.) and all words that occur in less than 1% of the reviews in the data set (Griffiths and Steyvers 2004, Büschken and Allenby 2016, Puranam et al. 2017, Berger et al. 2020). After preprocessing, there are 785 unique words in the vocabulary, and the average length of the textual reviews is 37 words (s.d. = 30.83). Note that this step is at the review level, so we preprocess all available 889,660 consumer reviews for the 262 awarded restaurants and 1,257 control restaurants.

The LDA model assumes a certain data-generating process for the review text: When consumers write reviews, they can choose words to express their opinions about multiple dimensions (i.e., topics) of the dining experience, such as food and service. Thus, each review $d \in \{1, ..., D\}$ includes a mixture of K topics, and each topic $k \in \{1, ..., K\}$ is characterized by a probability distribution over a vocabulary of V words $v \in \{1, ..., V\}$. The standard LDA model assumes the same Dirichlet prior for all of the per-review topic distributions (α) and the same prior for all of the per-topic word distribution (β). In other words, it ignores review and word characteristics that might affect the distribution of topics. To account for potential heterogeneity, we allow the hyperparameter α_d to be a function of the length and rating of the textual review. If two reviews have few characteristics in common, their Dirichlet prior α_d should be different, resulting in the different topic distribution $\boldsymbol{\theta}_d$. For instance, a longer review might discuss more topics and therefore have more evenly spread topic distributions. Similarly, we allow the hyperparameter β_k to be a function of latent semantic word characteristics. If two words have different semantic characteristics (e.g., they are antonyms rather than synonyms), we expect

that these two words will have different probabilities of appearing in the same topic k. In other words, if a topic "prefers" a certain word v, we expect that it will also prefer other words with similar semantic characteristics to v.

At the review level, we segment the review characteristics into quintiles based on the length of the review (measured by the number of words before preprocessing) and its rating. Thus, the review characteristics are represented by two categorical 5-level variables, which are further converted into L_{doc} dummy variables, each corresponding to one level of the $5 \times 5 = 25$ combinations of the review length quintile and the star rating, so $L_{doc} = 25$ in our model. In addition to the observable review characteristics, we add an intercept term to capture the characteristics that are unrelated to the L_{doc} binary variables. Therefore, the characteristics of review d are defined by an (L_{doc}+1)-dimensional binary vector $\boldsymbol{f_d} = \{1, f_{d,1}, f_{d,2}, ..., f_{d,l}, ..., f_{d,L_{doc}}\}^T$, where $f_{d,l}$ equals 1 if review d has the characteristic indicated by label l and 0 otherwise (Zhao et al. 2017).

At the word level, frequency counts of word occurrences in a corpus are the primary data to all unsupervised methods for learning word representations. However, standard LDA approaches do not consider word characteristics, presenting challenges with short texts, where word co-occurrences are too sparse to provide meaningful context. For example, it is possible that topics associated with synonyms have a prior tendency to be similar, so that when one synonym is rare but the other is common within the corpus, the topics estimates for the rare one can be improved. A global log-bilinear regression model GloVe provides an effective measure for the linguistic or semantic similarity of word representations (Pennington et al. 2014). Under GloVe representations, each word is represented by a high dimensional vector that is pre-trained on some large external corpus, e.g., Wikipedia, Twitter, and Google News. Accordingly, we choose a set of 50-dimensional word

embeddings pre-trained on Twitter[18] as our original word characteristics. Similar to the review characteristics, we convert the continuous-valued word characteristics into binary values, following Guo et al. (2014). Let $\boldsymbol{M''}$ be a $V \times 50$ matrix, where V is our vocabulary size. Each row $v \in \{1, ..., V\}$ of $\boldsymbol{M''}$ represents a 50-dimensional embedding of vocabulary word v. For the j^{th} dimension ($j \in \{1, ..., 50\}$) of word embeddings, we divide the corresponding column vector $\boldsymbol{M''_{.j}}$ into two parts, with one part including all positive elements ($M''_{.j+}$) and the other including the negative elements ($\boldsymbol{M''_{.j-}}$). Next, we transform $\boldsymbol{M''}$ to a same-dimension matrix $\boldsymbol{M'}$ as follows:

$$M'_{v,j} = \begin{cases} 1 & if\ M''_{v,j} > mean(\boldsymbol{M''_{.j+}}), \\ -1 & if\ M''_{v,j} < mean(\boldsymbol{M''_{.j-}}) \\ 0 & otherwise \end{cases},$$

where mean($\cdot$) denotes the mathematical mean. The insight behind this transformation is that we only consider the word embeddings with strong positive or negative values on each dimension j and omit the values that are close to 0. Finally, we use two dummy variables to encode each column j in $\boldsymbol{M'}$ and transform $\boldsymbol{M'_{v,j}} \in \{-1, 0, 1\}$ to binarized word characteristics. Thus, the original continuous-valued word labels are converted to L_{word} unique binary labels ($L_{word} = 100$ in this case). The labels of each word $v \in \{1, ..., V\}$ are defined by an (L_{word}+1)-dimensional binary vector $\boldsymbol{g_v} = \{1, g_{v,1}, g_{v,2}, ..., g_{v,l'}, ..., g_{v,L_{word}}\}^T$, where $g_{v,l'}$ equals 1 if word v has the characteristic indicated by label l' and equals 0 otherwise.

The LDA model describes the joint probability distribution over both the observable data (words in the review) and the hidden variables (topics of the review). In our LDA model, we allow the Dirichlet prior $\boldsymbol{\alpha}_d$ to be a function of review characteristics $\boldsymbol{f_d}$, and the Dirichlet prior $\boldsymbol{\beta}_k$ to be a function of word characteristics $\boldsymbol{g_v}$,

[18]The word embedding was pre-trained on 2 billion tweets with 1.2 million unique words by Pennington et al. (2014).

specified as follows

$$\alpha_{d,k} = exp(\sum_{l=1}^{L_{doc}} f_{d,l}\, \lambda_{l,k}) = exp(\boldsymbol{f_d}^T \boldsymbol{\lambda_k}), \quad \lambda_{l,k} \sim F(\alpha_{d,k}) \tag{2.4.1}$$

$$\beta_{k,v} = exp(\sum_{l'=1}^{Lword} g_{v,l'}\, \delta_{l',k}) = exp(\boldsymbol{g_v}^T \boldsymbol{\delta_k}), \quad \delta_{l',k} \sim G(\beta_{k,v}) \tag{2.4.2}$$

where F(ů) and G(ů) denote a function of parameters inside (Zhao et al. 2017). We initialize the value of $\alpha_{d,k}$ as $1/K$, i.e., equal probability for K topics per review. After $\lambda_{l,k}$ is sampled, we can update the value of $\alpha_{d,k}$ and iterate over the (L_{doc}+1)-dimensional vector $\boldsymbol{f_d}$. Similarly, we initialize the value of $\beta_{k,v}$ as 0.01 (i.e., equal probability for 100 words per topic), and update $\beta_{k,v}$ by iterating over the (L_{word}+1)-dimensional vector $\boldsymbol{g_v}$.

We vary the number of topics between two and ten,[19] and estimate the LDA model incorporating both review-level and word-level characteristics by Markov Chain Monte Carlo (MCMC).[20] We find that the LDA model with five topics yields the highest topic coherence score, a measurement that has been shown to make the resulting topics more interpretable (Chang et al. 2009, Röder et al. 2015, Zhang and Luo 2023). We therefore set the number of topics $K = 5$, and estimate the LDA model with both review-level and word-level characteristics.

Table 2.8 displays the top 20 words in descending order in terms of the posterior probability to be associated with each topic. It appears that Topic 5 discusses the general dining experience with an overall evaluation, whereas Topics 1–4 discuss the dining experience in four different respects. Topic 1 concerns perceptions of cost-effectiveness, and it relates to value for money. Generally, consumers think

[19] A larger number of topics is less preferred because it produces topics with a lot of overlap.

[20] We maximize the likelihood of the topic assignments for each word in the corpus with respect to the parameters $\lambda_{l,k}$ and $\delta_{l',k}$, and obtain the review-level topic proportions $\boldsymbol{\theta}_d$. We run the MCMC chain for 15,000 iterations, with the first 1,500 iterations as burn-in. The hyperparameters $\boldsymbol{\alpha}_d$ and $\boldsymbol{\beta}_k$ are estimated and optimized every 100 iterations.

the experience is good but might be overpriced, as evidenced by the words "price," "bite," and "expensive." Topic 3 centers around the menu and food, as evidenced by the use of words such as "starter," "dessert," "steak," "beef," "fish," "cheese," etc. Topic 5 includes words describing the general experience in various aspects (e.g., "experience," "wine," "food," "dining," "meal"). Both Topic 2 and Topic 4 relate to service but are associated with different valence. Topic 2 relates to complaints about services, such as issues regarding time ("time," "wait," "minute"), as well as interactions with service personnel ("waiter," "staff," "ask," "come"), which possibly relate to attempts to resolve issues.[21] Topic 4 relates to positive service encounters, because most adjectives for this topic have a positive valence ("great," "excellent," "amazing," "friendly," and "attentive"). Our descriptions of topics continue to hold when only considering words that are unique to each of the five topics (see Table 2.26 in Appendix 2.8.3).

We verify our interpretation of topic valence by checking the correlations between a review's overall rating and the probability of being associated with each of the five topics (value for money, issues with order, menu and food, service and staff, and overall experience).[22] In Table 2.9, Topic 5 (overall experience) is positively

[21] As an example, a representative review that has a high probability ($\theta > 0.85$) for Topic 2 (issues with order) is presented below. It was posted after the restaurant gained Michelin stars, which provides further evidence that consumers might have higher expectations after a restaurant gains Michelin stars. *"Been several times prior to the changes and the Michelin star award so maybe expectations were too high. On arrival seated ourselves in the bar, staff were busy in and out of restaurant no welcome smile or will be with you soon. Totally ignored. After about 10 minutes someone came to take drinks order was very pleasant and hospitable. Nice table taken to on time, extremely disappointed to be told on seated that there was only one lamb left which we immediately reserved. On taking our order we did politely express our disappointment that of only two meat choices one was not available, the response from the waitress was a shrug and well they are closed for the next two days! One of our party of 4 was very disappointed with the roast potatoes, tasted not fresh but rather as if been keep warm for hours. When paying the bill, a very reasonable bill for Michelin star, we did raise our complaints they were not received very well. Whatever business one is in, how complaints are treated gives an insight on the company and their standards, flitch of bacon came up wanting in this area more than in any other. Poor defensive excuses of new staff not properly trained, well they should have been."*

[22] The topic probabilities for each review add up to one.

correlated with the overall rating, so the higher Topic 5 probability, the higher review ratings. We find a correlation of –0.521 ($p < 0.0001$) between Topic 1 (value for money) and the overall rating. This is intuitive: consumers may be more likely to mention value for money when it is low, which may make them less satisfied. Surprisingly, Topic 3 (menu and food) has a negative correlation of –0.174 ($p < 0.0001$) with the overall rating, possibly because consumers tend to complain about food when mentioning it. While both Topic 2 and Topic 4 relate to service, Topic 2 (issues with order) is negatively correlated with the overall rating, whereas Topic 4 (service and staff) is positively correlated with the overall rating. These correlations are consistent with our topic interpretation above.

Table 2.8: Top 20 Words Under the LDA Model (K = 5)

Rank	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
	Value for Money	Issues with Order	Menu and Food	Service and Staff	Overall Experience
1	food	table	main	food	menu
2	good	ask	starter	service	course
3	service	take	dessert	great	wine
4	price	order	cook	staff	experience
5	place	time	steak	excellent	food
6	great	get	good	recommend	dish
7	nice	book	dish	visit	tasting
8	menu	arrive	course	lovely	well
9	really	drink	delicious	amazing	star
10	wine	come	beef	friendly	every
11	bite	tea	meal	good	dining
12	get	staff	order	place	meal
13	like	wait	fish	time	chef
14	expect	say	cheese	lunch	eat
15	better	waiter	taste	delicious	staff
16	quality	minute	sauce	atmosphere	time
17	staff	bar	chocolate	definitely	visit
18	little	service	bread	attentive	win
19	expensive	leave	serve	birthday	michelin
20	quite	tell	menu	fantastic	best

Table 2.9: Correlations Between Overall Review Rating and Topic Probabilities

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
	Value for Money	Issues with Order	Menu and Food	Service and Staff	Overall Experience
Overall Rating	–0.521***	–0.551***	–0.174***	0.261***	0.513***

Note: *** p<0.01.

Effects of Michelin Stars on Topics of Consumer Reviews

Given the topic distributions obtained from the LDA model, we aggregate the review-level topic distribution to restaurant level in the 90-day guidebook windows for SCM-DiD and in the two-month blocks for SynthDiD. Then, we analyze the effect of Michelin star changes on the topics of consumer reviews with models described in Section 2.3, using as dependent variable the mean probability of each of the five topics. Tables 2.10 and 2.11 respectively report the SCM-DiD and SynthDiD estimation results.

Column (5) of Table 2.10 shows that a decrease in Michelin stars is associated with an increase in the discussion of overall experience (Topic 5). As overall experience (Topic 5) is positively correlated with the review's overall rating, the result is consistent with our prior findings on sentiment of consumer reviews. We find that consumers are more likely to discuss value for money (Topic 1, Column (1)) when a restaurant gains Michelin stars, and are less concerned about it when a restaurant loses Michelin stars. This is consistent with reference dependence (Gerstner 1985, Winer 1986, Rao and Monroe 1989, Almenberg and Dreber 2011) and expectation effect: consumers raise their expectations and become more demanding with recommendations from experts. Regarding menu and food (Topic 3, Column (3)), we note that consumers tend to mention these aspects less frequently when a restaurant loses Michelin stars, possibly because they have lower expectations about food in such cases. Finally, for the two service-related topics (Topic 2 and Topic 4), an increase in Michelin stars is associated with an 8.8 percentage point increase in the proportion

of Topic 2 and a 17.4 percentage point decrease in the proportion of Topic 4. In contrast, a decrease in Michelin stars is associated with a 9.2 percentage point decrease in the proportion of Topic 2 and a 17.5 percentage point increase in the proportion of Topic 4. This suggests that consumers become more demanding on service quality when restaurants gain Michelin stars, and less demanding when restaurants lose Michelin stars. Note that the results are not driven by menu changes, because we focus on restaurants without menu changes in the guidebook windows. The results from the SynthDiD estimation in Table 2.11 show a similar pattern.

Together, our results on the content of consumer reviews shed light on the underlying factors behind the changes in review sentiment following changes in Michelin stars, providing support on the expectation effect. Service and "value for money" are crucial to customer satisfaction. This finding is highly relevant to practitioners as they navigate the impacts of expert opinions. As our results show, receiving a favorable expert opinion can put more pressure on the business due to raised customer expectations. Thus, practitioners need to be proactive in terms of anticipating this and preparing accordingly.

Table 2.10: Effects of Michelin Star Changes on Topics of Consumer Reviews by SCM-DiD

	(1) Topic 1 Value for Money	(2) Topic 2 Issues with Order	(3) Topic 3 Menu and Food	(4) Topic 4 Service and Staff	(5) Topic 5 Overall Experience
After	0.006	–0.004	0.000	–0.003	0.001
	(0.004)	(0.003)	(0.004)	(0.005)	(0.004)
After × Increase	0.131***	0.088***	–0.018	–0.174***	–0.027
	(0.021)	(0.015)	(0.011)	(0.018)	(0.025)
After × Decrease	–0.160***	–0.092***	–0.099***	0.175***	0.175***
	(0.016)	(0.014)	(0.012)	(0.023)	(0.024)
One Star	0.004	–0.001	–0.006	–0.002	0.005
	(0.019)	(0.014)	(0.011)	(0.019)	(0.025)
Two Star	–0.016	–0.008	–0.021	0.042	0.002
	(0.041)	(0.026)	(0.022)	(0.031)	(0.054)
Three Star	–0.133	–0.054	0.032	0.144***	0.010
	(0.125)	(0.086)	(0.025)	(0.046)	(0.216)
ln(number of reviews+1)	0.003	0.010**	0.009	–0.009	–0.013
	(0.009)	(0.005)	(0.008)	(0.011)	(0.012)
Cumulative average rating	–0.034	–0.067	–0.070	0.010	0.162
	(0.098)	(0.076)	(0.078)	(0.099)	(0.126)
Cumulative rating variance	0.014	–0.036	–0.132**	0.059	0.095
	(0.075)	(0.069)	(0.063)	(0.065)	(0.071)
ln(normalized search volume+1)	–0.023	0.012	–0.029*	0.014	0.027
	(0.027)	(0.022)	(0.016)	(0.042)	(0.041)
Pair–window FE	Yes	Yes	Yes	Yes	Yes
Restaurant FE	Yes	Yes	Yes	Yes	Yes
Observations	892	892	892	892	892
Number of pairs	223	223	223	223	223
R^2	0.861	0.827	0.785	0.885	0.901

Note: Robust standard errors clustered at pair level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 2.11: Effects of Michelin Star Changes on Topics of Consumer Reviews by SynthDiD

	(1) Topic 1 Value for Money	(2) Topic 2 Issues with Order	(3) Topic 3 Menu and Food	(4) Topic 4 Service and Staff	(5) Topic 5 Overall Experience
Increase	0.071***	0.038***	–0.015	–0.111***	0.018
	(0.021)	(0.013)	(0.012)	(0.020)	(0.017)
Decrease	–0.063***	–0.054**	–0.044***	0.144***	0.105***
	(0.024)	(0.024)	(0.017)	(0.039)	(0.030)

Note: Overall ATT reported. Aggregated standard errors are in parentheses. *** p<0.01, ** p<0.05.

2.5 Alternative Explanations

Although two variants of the synthetic control method allow us to capture possible trends that might affect identification of the effect of the Michelin star change, as summarized in Table 2.4, there still exist potential supply- and demand-side factors that may lead to the observed effects. We address concerns related to supply-side factors in Section 2.5.1 and concerns related to demand-side factors in Section 2.5.2.

2.5.1 Supply-side Factors

There are three supply-side changes that may affect consumer reviews: menu changes, changes in serving size given the menu, and changes in restaurant decor and/or service. We discuss each in turn.

Menu Changes

One alternative explanation to the finding is that restaurants may have changed their menu following the Michelin star change. Recall that our sample excludes restaurants that have changed their menus during the window around the Michelin Guide release time, thus it is unlikely that the effects are driven by menu changes.

Table 2.12: Subset of Restaurants Evidencing Consistency

	(1) SCM-DiD	(2) SynthDiD
Increase	0.001	0.010
	(0.074)	(0.069)
Decrease	0.340***	0.351***
	(0.079)	(0.155)

Note: Regression coefficients on Equation (2.3.1) are reported in Column (1), and overall ATTs estimated by Equations (2.3.2) and (2.3.3) are reported in Column (2). Robust standard errors clustered at pair level (Column 1) and aggregated standard errors (Column 2) are in parentheses. *** p<0.01.

Serving Size Changes

Although we have controlled for menu offerings and focused on a short-time window, one concern is that restaurants can modify serving sizes or the quality of their dishes without changing the menu. As discussed in Section 2.2.1, the Michelin star selections are confirmed through repeated visits by different inspectors within a year, ensuring consistency. Should there be changes in serving size or food quality post a Michelin star status change, it would likely be noted by the inspectors during their consistency assessments and could result in an adjusted star rating the following year. Thus, restaurants that retain their new Michelin star level in the next guidebook year (e.g., sustaining a 1-star status after an increase from 0-star) are presumed to uphold consistent food quality and serving sizes. We replicate the analysis with this subset of "highly consistent" restaurants, and the results in Table 2.12 are consistent with prior results.[23][24]

Non-food Changes

The third concern related to the supply-side is that restaurants may have made major changes in their decor or service. In SCM-DiD, we focus on a window of 90 days before and 90 days after the Michelin Guide release, a timeframe likely too short for significant changes. We revisit this issue by conducting a robustness check with a shorter window period in Section 2.6.3. In SynthDiD, Figure 2.4 indicates a very significant positive treatment effect in an even shorter period (60 days) after the new Michelin Guide release. This makes it even less likely that changes in decor and service could be causing our results.

[23]Note that our data period ends at guidebook 2020 and does not cover Michelin star levels in guidebook 2021, thus this analysis includes star changes before guidebook year 2020.

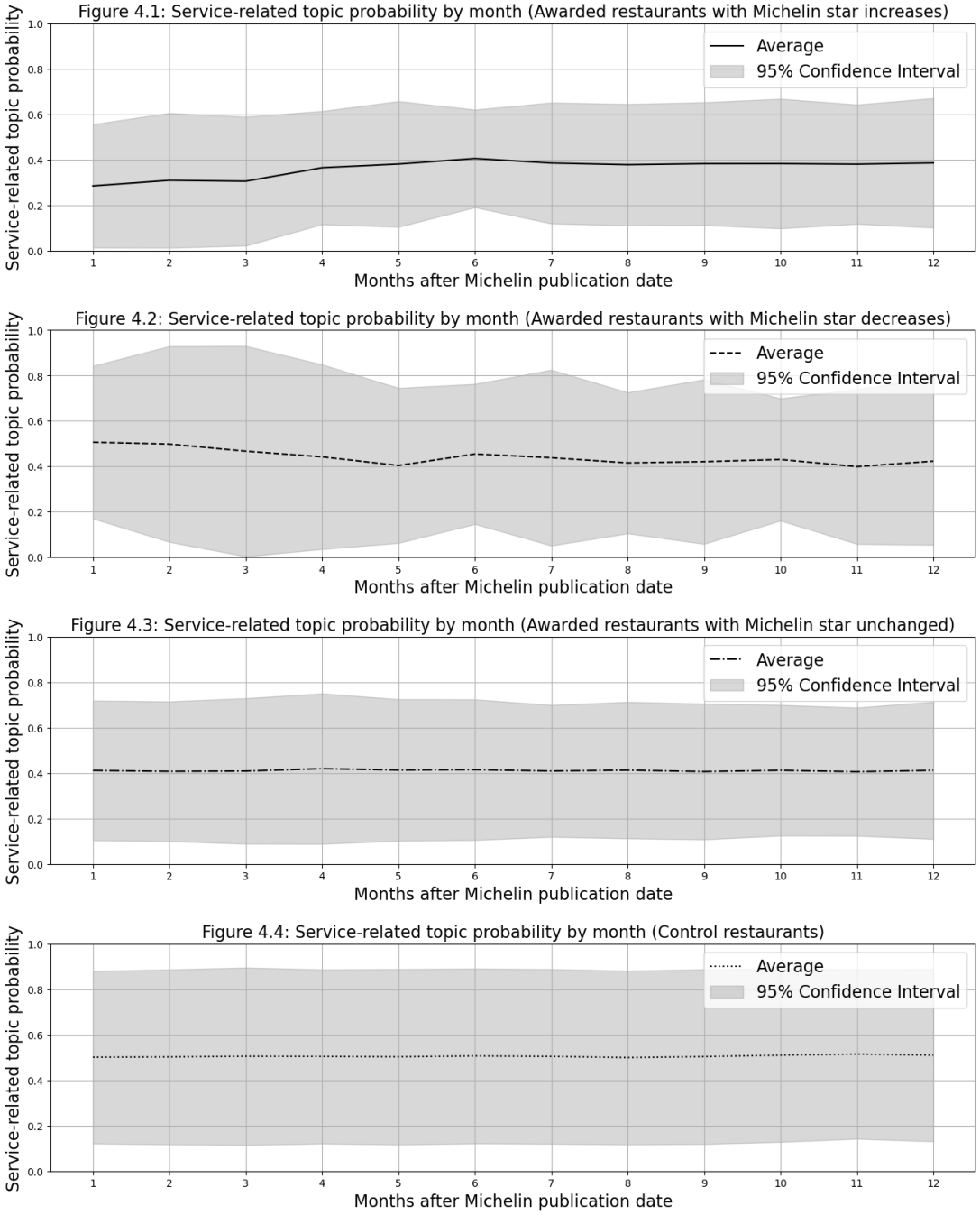
[24]Note that, for brevity, in Table 2.12 and subsequent tables, we do not report the estimates associated with other control variables in SCM-DiD, which are qualitatively similar to those in Column (2) of Table 2.6.

To further rule out the possibility that service changes drive the effect, we analyze trends in review topics that might reflect such changes following the Michelin Guide updates. To do so, we focus on "service-related" metrics, based on topics 2 and 4 (cf. Section 2.4.2). This involves aggregating the probabilities of relevant topics over the twelve month period between guidebook releases. We follow a three-step procedure: First, we categorize all the restaurants by guidebook years into four groups: awarded restaurants with Michelin star increases (169 units); awarded restaurants with Michelin star decreases (83 units); awarded restaurants whose Michelin star status remained unchanged (2,091 units); and control restaurants (7,517 units). Second, for each unit within each of these four restaurant groups, we aggregate the reviews by month following the Michelin Guide publication date. Third, for each of these four restaurant groups, we plot the average service-related topic probability, aggregated across restaurants and Michelin Guidebook years, along with their 95% confidence intervals, as shown in Figure 2.5. This additional analysis expands the period in our previous analyses from a maximum of 90 days to a full year.

If the star changes led to adjustments in service levels, we should expect to observe a corresponding shift in the probability of this topic being mentioned for restaurants that experienced a Michelin star change. However, as shown in Figure 2.5, for all four groups, the probability of service-related topics being mentioned remains stable over the twelve month period. The trends in the third and fourth groups are more stable, because of the substantially larger numbers of observations.

This being said, we acknowledge that with sufficient commitment from management, there is a possibility of relatively swift improvements in service quality. Because consumer reviews reflect both objective service quality and subjective perceptions influenced by expectations, we cannot completely rule out the potential impact of unobserved service quality adjustments. However, based on the above discussion and analyses, it seems unlikely that these are the primary drivers of the observed effects.

Figure 2.5: Service-related Topic Probability by Month



Note: To plot this figure, we follow a three-step procedure. First, we categorize all the restaurants by guidebook years into four groups: awarded restaurants with Michelin star increases (169 units); awarded restaurants with Michelin star decreases (83 units); awarded restaurants whose Michelin star status remained unchanged (2,091 units); and control restaurants (7,517 units). Second, for each unit within each of these four restaurant groups, we aggregate the reviews by month following the Michelin Guide publication date. Third, for each of these four restaurant groups, we plot the average service-related topic probability, aggregated across restaurants and Michelin Guidebook years, along with their their 95% confidence intervals.

2.5.2 Demand-side Factors

There are three demand-side changes that may affect the interpretation of our results: changes in restaurant demand, consumers showing sympathy for restaurants losing stars, and changes in the mix of consumers visiting a restaurant. We discuss each in turn.

Restaurant Demand

Michelin star changes may induce changes in consumer interest and restaurant demand. To see how Michelin stars affect restaurant demand, we estimate Equations (2.3.1), (2.3.2) and (2.3.3) with the log-transformed normalized Google search intensity (collected from Google Trends) as the dependent variable (see Table 2.13). Our findings reveal that changes in a restaurants Michelin star status do not significantly change its search volume, suggesting that our results are unlikely to be primarily driven by changes in restaurant demand.

Table 2.13: Google Trends Search Volume

	DV: log-transformed normalized Google search intensity	
	(1) SCM-DiD	(2) SynthDiD
Increase	0.019	0.022
	(0.064)	(0.071)
Decrease	0.004	–0.082
	(0.059)	(0.088)

Note: Regression coefficients on Equation (2.3.1) are reported in Column (1), and overall ATTs estimated by Equations (2.3.2) and (2.3.3) are reported in Column (2). Robust standard errors clustered at pair level (Column 1) and aggregated standard errors (Column 2) are in parentheses.

This being said, Google search intensity only includes searches originated from

Google, and it is possible that changes in Michelin stars lead to changes in searches on review and booking websites such as TripAdvisor and OpenTable. We next examine restaurant demand using daily OpenTable reservation data collected by Farronato and Zervas (2022) on New York City restaurants. This dataset contains information on the daily availability of tables for two between 18:30 and 19:30 at each restaurant in the period of April 2013 to March 2017. Within this time period, we first check five New York City Michelin Guides (guidebook 2013 to guidebook 2017), and identify 117 awarded restaurants that received Michelin stars at least once. Among these awarded restaurants, there are 54 instances of Michelin star increase and 39 instances of star decrease during guidebook 2014 to guidebook 2017 (with guidebook 2013 serving as our baseline). Second, we match these awarded restaurants with the restaurants in Farronato and Zervas (2022)'s OpenTable reservation data, and identify 70 (out of 117) awarded restaurants with OpenTable records. Third, we denote each "restaurant-guidebook year" as a unit, and keep units for which we observe booking information immediately before and after the guidebook release. In the end, we retain a total number of 222 units, with 27 units associated with Michelin star increases, 13 units associated with Michelin star decreases, and 182 units where star status remained unchanged. We then estimate the following regression model analogous to Equation (8) from Farronato and Zervas (2022), using data in a short window around the guidebook release dates:

$$\begin{aligned} Soldout_{id} = \quad & \beta_1 After_d + \beta_2 After_d \times Increase_{id} + \beta_3 After_d \times Decrease_{id} \\ & + \beta_4 OneStar_{id} + \beta_5 TwoStar_{id} + \beta_6 ThreeStar_{id} \\ & + \alpha_i + \gamma_d + \varepsilon_{id} \end{aligned} \tag{2.5.1}$$

where i denotes restaurant, and d denotes day. The outcome variable $Soldout_{id}$ is an indicator variable which equals 1 if restaurant i is sold out between 18:30 and

19:30 on day d. $After_d$ is an indicator variable which takes the value of 1 if day d is in a window after the Michelin guidebook update. $Increase_{id}$ ($Decrease_{id}$) takes the value of 1 if restaurant i gained (lost) stars in the corresponding new guidebook. We control for restaurant fixed effect α_i and day fixed effects γ_d. Table 2.14 shows the results. Column 1 (2) is based on a window of 60 (90) days before and 60 (90) days after the guidebook release dates.

The results indicate that compared to restaurants that maintained the same Michelin star level, restaurants gaining Michelin star(s) experience an increase in demand, whereas those losing Michelin star(s) do not experience a significant change in demand. Although this sample of restaurants differs from our main sample, we posit that the relationship between Michelin stars and restaurant demand applies generally. This implies that the observed effects of Michelin star decreases on consumer reviews (Section 2.4) are unlikely to be driven by changes in restaurant demand. For restaurants gaining Michelin star(s), we conjecture that an increase in demand could potentially compromise the dining experience (possibly due to overcrowding etc.). The fact that the consumer review ratings in our main sample remained stable despite increased demand further suggests that the results are unlikely driven by changes in restaurant demand.

Consumer Sympathy

One alternative explanation for our results is that consumers show their sympathy to underdogs (i.e., restaurants losing Michelin stars) and thus try to defend them in reviews. If this were the main mechanism, we would expect an increase in review volume for restaurants losing Michelin stars. To test if this is the case, we estimate Equations (2.3.1), (2.3.2) and (2.3.3) with the volume of consumer reviews as the dependent variable (Table 2.15). We do not find significant changes in review vol-

Table 2.14: Effect of Michelin Stars Changes on Restaurant Demand (New York City)

	DV: whether 18.30 – 19.30 slot sold out on OpenTable	
	60-day window	90-day window
	(1)	(2)
After	0.040***	0.063***
	(0.015)	(0.015)
After × Increase	0.084**	0.078**
	(0.034)	(0.031)
After × Decrease	–0.014	–0.026
	(0.047)	(0.039)
One Star	–0.056	–0.046
	(0.041)	(0.036)
Two Star	0.191***	0.212***
	(0.071)	(0.058)
Three Star	0.209***	0.227***
	(0.075)	(0.058)
Restaurant FE	Yes	Yes
Day FE	Yes	Yes
Observations	24,987	35,375
Number of units	222	222
R^2	0.460	0.455

Note: Robust standard errors clustered at restaurant level are in parentheses.
*** p<0.01,** p<0.05.

umes for restaurants gaining or losing Michelin stars,[25] suggesting our results are unlikely an outcome of the consumer sympathy to underdogs.

Consumer Mix

Potential changes in Michelin stars might change the mix of consumers who visit the restaurant. There are two possible mechanisms that could lead to the change in customer experience.

First, a change in the Michelin star ratings does not change the consumer mix

[25]Consumer sympathy could potentially be more evident for British cuisine restaurants, as consumers might be inclined to support their national cuisine. To explore this, we replicate our analysis with restaurants serving British cuisine and the results are consistent. The results are available upon request.

Table 2.15: Volume of Consumer Reviews

	DV: review volume	
	(1) SCM-DiD	(2) SynthDiD
Increase	8.042	0.735
	(5.426)	(1.977)
Decrease	–7.971	–2.412
	(4.939)	(2.996)

Note: Regression coefficients on Equation (2.3.1) are reported in Column (1), and overall ATTs estimated by Equations (2.3.2) and (2.3.3) are reported in Column (2). Robust standard errors clustered at pair level (Column 1) and aggregated standard errors (Column 2) are in parentheses.

visiting the restaurant. Thus any change in experience is driven by the change in expectations. Second, a change in the Michelin star ratings does change the consumer mix visiting the restaurant. Thus any change in experience is driven by a combination of selection and the change in expectations.

In either case, what the restaurant cares about is the aggregate level experience presented to interested consumers. In other words, going forward, a prospective (or even repeat) diner is unlikely to think about the (changing) consumer mix in her decision to visit the restaurant. Therefore, the potential change in consumer mix is unlikely to affect the implications of the findings for the restaurant, at least in the short to medium term.

Having said this, it is important to provide evidence that the second mechanism (above) is unlikely to be at play. To do this, we use the reviewer-level data described in Section 2.2.6 for two sets of analyses. First, we look at the characteristics of all the reviewers who have reviewed the focal restaurant *before* and the characteristics of those who reviewed *after* the star change. Second, we look at the reviewers of the focal restaurant and examine their behavior in terms of the characteristics of *all* the restaurants i.e., not just the ones in our sample, that they visit before and after the star change.

Restaurant-level Analysis: Reviewer Characteristics As described in Section 2.2.6, we have collected the TripAdvisor profiles of 52,210 unique reviewers, who have provided 1,617,923 reviews spanning from 2010 to 2020, of which 52,224 reviews are for awarded restaurants. Based on this dataset, we construct a series of reviewer-level characteristics. We then compare characteristics of reviewers who reviewed the restaurant *before* the Michelin star change against those who reviewed *after* the change. If we observed no significant changes in these characteristics, it would provide us with greater confidence that the change in consumer mix is not the main driver behind our findings. We detail the steps below.

First, we introduce four variables to describe reviewer characteristics based on their profile, as shown in Table 2.16. To illustrate the construction of these variables, consider a reviewer who is registered in the United States and has provided eight reviews. Among the eight reviews sorted in chronological order, the fifth and eighth reviews are for two awarded restaurants, each receiving a "5-star" rating. The remaining six reviews have "4-star" ratings. The "Example" column shows the values of the four variables for this reviewer. The variable "Local consumer" takes the value of 0 because she is not registered in the United Kingdom or Ireland. The variable "Picky consumer" takes the value of 0 because she has given 5-star review ratings. The cumulative number of restaurants until the two awarded restaurants are respectively 4 and 7. The cumulative mean review rating until the first awarded restaurant is 4 because the previous four reviews all have 4-star ratings. The cumulative mean review rating until the second awarded restaurant is 4.14 because among the seven previous reviews, six have 4-star ratings and one has a 5-star rating.

Next, for each of these 52,224 reviews, we extract the reviewer's characteristics at the time of the review. For illustration, Table 2.27 in Appendix 2.8.4 shows the reviewer characteristics associated with the two reviews by the example reviewer presented in Table 2.16.

Lastly, similar to the data preparation step of SCM-DiD, we aggregate the re-

Table 2.16: Reviewer Characteristics

Variable	Definition	Example
Local consumer	Equals 1 if the reviewer is registered in the "United Kingdom" or "Ireland". Otherwise, equals 0.	0
Picky consumer	Equals 1 if the reviewer has never given a "5-star rating" in their profile. Otherwise, equals 0.	0
Cum. # of restaurants until each awarded restaurant	The number of restaurants that a reviewer has reviewed until each awarded restaurant.	4 and 7
Cum. mean review rating until each awarded restaurant	Mean review rating across all previously reviewed restaurants.	4 and 4.14

views at the restaurant level for both the pre- and post- windows, and then use the mean consumer review rating (or the mean topic probability) as the dependent variable. In this specific analysis, we aggregate the reviewer characteristics constructed at the review level – the four variables listed in Table 2.27 – at the restaurant level. The resulting average of the review-level reviewer characteristics for each restaurant is the dependent variable. Essentially, when aggregating the variables "local consumer" and "picky consumer" at the restaurant level, we are measuring the percentage of "local (picky) consumers" associated with the restaurant. When aggregating the other two variables at the restaurant level, we are measuring the mean value of those variables (averaged across reviewers) for the restaurant.

In this analysis, we have 217 awarded restaurants and 1,040 "restaurant-guidebook year" units. Among these 1,040 units, 71 are associated with Michelin star increases, 53 are associated with Michelin star decreases, and the remaining units represent cases where the Michelin star status remained unchanged. Each of these units comprises two observations: one for the period before the Michelin star rating change and the other for the period after the Michelin star rating change.

The results of the analysis are shown in Table 2.17. Columns (1) and (2) indicate that the percentage of local consumers and the percentage of picky consumers do not change significantly after Michelin star changes. Columns (3) and (4) suggest that the review intensity and average review rating are similar between reviewers who

reviewed the restaurant prior to the change and those who reviewed it afterwards. Together, these results suggest that Michelin star changes do not have a significant impact on the types of consumers who visit awarded restaurants. Consequently, it implies that a change in the consumer mix is unlikely to be the primary driving factor behind our findings.

Table 2.17: Effect of Michelin Star Changes on Reviewer Characteristics (90-day Window)

	Percentage (restaurant-level)		Mean (restaurant-level)	
	Local Consumer (reviewer-level)	Picky Consumer (reviewer-level)	# of restaurants until each awarded restaurant (reviewer-level)	Mean review rating until each awarded restaurant (reviewer-level)
	(1)	(2)	(3)	(4)
After	0.044***	–0.001	1.355**	–0.009
	(0.008)	(0.004)	(.548)	(0.016)
After × Increase	0.027	–0.001	–2.301	0.015
	(0.022)	(0.010)	(1.719)	(0.042)
After × Decrease	–0.024	–0.005	–0.916	0.060
	(0.033)	(0.014)	(2.363)	(0.054)
Other control variables (Michelin stars, the number of reviews, cumulative average review rating, cumulative rating variance, Google search volume)	Yes	Yes	Yes	Yes
Restaurant FE	Yes	Yes	Yes	Yes
Guidebook year FE	Yes	Yes	Yes	Yes
Observations	2,080	2,080	2,080	2,080
"restaurant-guidebook year" units	1,040	1,040	1,040	1,040
Number of restaurants	217	217	217	217
R^2	0.040	0.016	0.278	0.021

Note: Robust standard errors clustered at restaurant level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Reviewer-level Analysis: Restaurant Characteristics We have shown that consumers who reviewed a restaurant before the Michelin change are not fundamentally different from consumers who reviewed the restaurant after the Michelin changes. Next, we examine whether changes in Michelin stars led consumers to visit different types of restaurants, drawing on all the reviews across every restaurant

that has been reviewed by a reviewer in our dataset.

As discussed in Section 2.2.6, we located the TripAdvisor pages for 279,359 (out of 327,852) restaurants that have been reviewed by 45,274 (out of 52,210) reviewers who have reviewed an awarded restaurant within the 90-day guidebook windows. In total, these reviewers have provided 1,101,305 reviews. For each of these reviews, we collect both time-invariant and time-varying characteristics of the corresponding restaurant at the time of the review. Specifically, the time-invariant characteristics, the price level and cuisine type, were extracted from the restaurant's Tripadvisor page. Then, we calculate the cumulative review characteristics (number of reviews, mean star rating, standard deviation of star rating) for the restaurant up to the review date, leveraging the dataset of 79 million reviews we have collected. We illustrate this process with an example in Table 2.28 in Appendix 2.8.4.

After computing restaurant characteristics at the time of each review, we aggregate these restaurant characteristics at the reviewer level. Specifically, for each review within a reviewer's profile, we calculate the cumulative restaurant characteristics of her previously reviewed restaurants. Again, we illustrate this process with an example in Table 2.29 in Appendix 2.8.4.

We then construct four variables to describe whether the restaurant is different from the reviewer's previously reviewed restaurants, as presented in Table 2.18. First, we check if the price level and cuisine type of the restaurant differ from those restaurants the reviewer had previously reviewed. Next, we look at whether the review rating stands out from the reviewer's previous ratings. To define what counts as "standing out", we look at whether the rating falls within a normal range, calculated as the average plus or minus one standard deviation ($mean \pm SD$). Lastly, we calculate the difference in ratings ($\Delta Rating$) by comparing the rating of the current review against the reviewers average rating up to that point.

To analyze whether a review in the reviewer's profile corresponds to a restaurant that differs from those she had reviewed previously, we use these four variables

Table 2.18: Reviewer-level Restaurant Characteristics Measurement and Definition

Variable	Definition
Whether new price level	Equals 1 if the restaurant has a different price level from those previously reviewed. Otherwise, equals 0.
Whether new cuisine type	Equals 1 if the restaurant has a different cuisine type from those previously reviewed. Otherwise, equals 0.
Whether rating out of range of $mean \pm SD$	Equals 1 if the review rating for the focal restaurant is out of the range of previous ratings ($mean \pm SD$). Otherwise, equals 0.
Rating difference, $\Delta Rating$	Difference between the focal review rating and the cumulative mean review rating

as dependent variables and estimate difference-in-differences models at the review level. The results are presented in Table 2.19. We control for both restaurant-level cumulative characteristics (e.g., average review rating, total number of ratings, and rating variance) and reviewer-level cumulative characteristics within their profile (e.g., average ratings, number of ratings, number of unique price levels, and number of unique cuisine types). We also add fixed effects on price level, cuisine type, reviewer, month, and guidebook year. Note that our analysis includes reviews starting from the third one in the reviewer's profile, because the initial two restaurants serve as a basis for computing rating variances and provide baseline price levels and cuisine types.

Table 2.19 shows that, across all four columns, there are no significant changes in restaurant characteristics at the reviewer-level after Michelin star changes. This suggests that consumers maintain their usual dining preferences, and thus, changes in Michelin stars do not appear to significantly influence consumers' decisions to visit these restaurants.

Overall, based on observables in a large amount of reviewer and review data, we find that the pool of diners at a focal restaurant does not change, and that reviewers of the focal restaurant do not exhibit any change in their restaurant choices/preferences, before and after Michelin star changes. This provides strong supportive

evidence that the second mechanism i.e., a change in consumer mix after a Michelin star change, is not driving our results.[26]

[26] An obvious caveat to this analysis is that we do not have data on restaurant visitors who do not write reviews at all (or write only on less prominent sites than TripAdvisor). Hopefully, the large sample sizes (in both analyses we carry out in this section) mitigate this concern.

Table 2.19: Effect of Michelin Star Changes on the Characteristics of Reviewed Restaurants

	Whether new price level	Whether new cuisine type	Whether rating out of range $mean \pm SD$	$\Delta Rating$
	(1)	(2)	(3)	(4)
After	0.002	0.006***	–2.150e-05	0.001
	(0.001)	(0.002)	(3.213e-04)	(0.001)
After × Increase	–0.019	–0.005	–2.225e-05	–0.001
	(0.013)	(0.009)	(4.206e-04)	(0.004)
After × Decrease	0.022	–0.013	–1.202e-04	–0.007
	(0.018)	(0.010)	(4.462e-04)	(0.006)
One Star	0.070***	0.004**	2.551e-04	0.001
	(0.003)	(0.002)	(2.579e-04)	(0.001)
Two Star	0.052***	0.018***	9.021e-04*	0.001
	(0.005)	(0.004)	(5.043e-04)	(0.002)
Three Star	0.014	0.050***	1.340e-03	–0.002
	(0.009)	(0.009)	(1.011e-03)	(0.004)
Cumulative average rating (Restaurant-level)	0.002***	0.001	–6.430e-03***	1.011***
	(0.001)	(0.001)	(4.816e-04)	(0.001)
Cumulative # of ratings (Restaurant-level)	–0.002**	0.004***	4.099e-04*	–0.002***
	(0.001)	(0.001)	(2.388e-04)	(0.001)
Cumulative rating variance (Restaurant-level)	–0.002*	0.003**	–6.253e-03***	–0.001
	(0.001)	(0.001)	(5.929e-04)	(0.001)
Cumulative average rating (Reviewer-level)	–0.010***	–0.003	–7.661e-04	–0.157***
	(0.002)	(0.002)	(6.790e-04)	(0.003)
Cumulative # of ratings (Reviewer-level)	–0.109***	–0.112***	–3.846e-03***	0.021***
	(0.005)	(0.006)	(4.103e-04)	(0.001)
Cumulative # of price levels (Reviewer-level)	0.057***	–0.005**	–4.817e-04**	–0.003***
	(0.002)	(0.002)	(1.953e-04)	(0.001)
Cumulative # of cuisine types (Reviewer-level)	0.001***	0.001***	3.900e-05***	–0.000**
	(0.000)	(0.000)	(1.150e-05)	(0.000)
Restaurant price level FE	Yes	Yes	Yes	Yes
Restaurant cuisine type FE	Yes	Yes	Yes	Yes
Reviewer FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Guidebook year FE	Yes	Yes	Yes	Yes
Observations	883,589	883,589	883,589	883,589
Number of reviewers	17,775	17,775	17,775	17,775
R^2	0.190	0.246	0.085	0.967

Note: Robust standard errors clustered at reviewer level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1

2.6 Robustness Checks

We conduct a battery of robustness checks, including a difference-in-differences analysis with control restaurants manually selected based on location, price, and cuisine type (Section 2.6.1), an alternative dependent variable to measure review sentiment (Section 2.6.2), an alternative window in SCM-DiD (Section 2.6.3), a falsification test with placebo guidebook publication dates in SynthDiD (Section 2.6.4), and a replication study with New York City data (Section 2.6.5).

2.6.1 Rule-based Control Restaurants

The SCM-DiD model in Section 2.3.2 employs the SCM to create a time-varying synthetic control restaurant that best matches the focal awarded restaurant, which is a data-driven approach. We check the robustness with a rule-based control restaurant selection, which explicitly selects control restaurants that closely match the awarded restaurants in terms of location, price level and cuisine type. Specifically, for each of the 262 awarded restaurants, we select from the pool of 1,147 control restaurants a control restaurant that satisfies the following criteria: (1) the control restaurant needs to be geographically close to the focal awarded restaurant: in urban areas within 0.5 miles and in rural areas within 10 minutes driving distance;[27] (2) the control restaurant has the same price level on TripAdvisor as the awarded restaurant; and (3) the control restaurant has the same cuisine type on TripAdvisor as the awarded restaurant. If more than one restaurant satisfies the above criteria, we give preference to the one that appears on the best nearby restaurants page recommended by TripAdvisor. Note that we allow each control restaurant to be matched with at most one awarded restaurant, that is we use matching without replacement to ensure that the results are not driven by a small group of control

[27]The distance is calculated by two restaurants' geocoded longitudes and latitudes. The travel time is estimated with Google Maps.

restaurants which are matched with many awarded restaurants. In the end, 227 (out of 1,147) control restaurants are selected, leading to 227 treated-control pairs. The remaining 35 awarded restaurants without identified control restaurants are either located in rural areas without nearby restaurants, or located in urban areas but do not have nearby restaurants with the same price level and cuisine type. Out of the 227 restaurant pairs identified, 156 are located in urban areas and 71 are located in rural areas. On average, the distance between the focal awarded restaurant and the selected paired control is 0.12 miles (s.d. = 0.29) in urban areas and 7.01 miles (s.d. = 8.48) in rural areas.

We then estimate the difference-in-differences model (Equation (2.3.1)) with the treated-control restaurant pairs where both restaurants have received reviews in the 90-day pre- and post-treatment windows. This results in 143 (out of 227) restaurant pairs. We report the results on review sentiment (Column 1) and review content (Columns 2-6) in Table 2.20. The results are qualitatively similar to those in Tables 2.6 and 2.10.

2.6.2 Alternative Sentiment Measure

One concern is that the proportion of extreme reviews (i.e., 5-star-rating and 1-star-rating reviews) has changed, but the mean review rating may not change. Following Shin et al. (2023), we replicate the review sentiment analysis with the percentage of 5-star-rating reviews at the restaurant level, instead of the mean review rating, as the outcome variable. Column (1) in Table 2.21 and Column (1) in Table 2.22 respectively replicate Column (2) in Table 2.6 (SCM-DiD) and Table 2.7 (SynthDiD). Results are consistent with our prior findings: decreases in Michelin stars improve consumer review ratings.

Table 2.20: Robustness Checks: DiD with Control Restaurants Selected via Rule-Based Criteria

	(1) Overall Rating	(2) Topic 1	(3) Topic 2	(4) Topic 3	(5) Topic 4	(6) Topic 5
		Value for Money	Issues with Order	Menu and Food	Service and Staff	Overall Experience
After	–0.097	0.014	0.011	–0.004	–0.025	0.004
	(0.064)	(0.012)	(0.012)	(0.010)	(0.019)	(0.010)
After × Increase	0.010	0.133***	0.066***	–0.032	–0.183***	0.016
	(0.112)	(0.032)	(0.024)	(0.022)	(0.039)	(0.036)
After × Decrease	0.216**	–0.171***	–0.086***	–0.079***	0.184***	0.153***
	(0.115)	(0.026)	(0.026)	(0.023)	(0.042)	(0.033)
Other control variables (Michelin stars, the number of reviews, cumulative average review rating, cumulative rating variance, Google search volume)	Yes	Yes	Yes	Yes	Yes	Yes
Pair–window FE	Yes	Yes	Yes	Yes	Yes	Yes
Restaurant FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	572	572	572	572	572	572
Number of pairs	143	143	143	143	143	143
R^2	0.708	0.811	0.699	0.737	0.801	0.900

Note: Robust standard errors clustered at pair level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

2.6.3 Alternative Window

As mentioned earlier, our main analysis with the restriction of the 90-day time window around the Michelin Guide release makes it hard for restaurants to have the time and/or resources to pull off major changes in decor and/or service levels. We further shorten the period to 60-day time window around the Michelin Guide release. Column (2) in Table 2.21 replicates Table 2.6 with a 60-day window in SCM-DiD,[28] and the results are robust.

2.6.4 Falsification Test

The SynthDiD model relaxes the strong parallel-trends assumption for all units and all time periods. However, it assumes that there exist unit and time weights such

[28]The re-construction of SCM results in 208 (out of 252) synthetic control restaurants corresponding to 208 treated units. 44 units were dropped because they do not have enough reviews on at least one side of the 60-day pre- or post-treatment window.

that the averaged treated unit and the weighted average of the control units satisfy a parallel trends assumption for the averaged post-treatment period and the weighted average of the pre-treatment periods (Arkhangelsky et al. 2021). In other words, the selection of weights on control units and pre-treatment periods depends on the actual treatment time. One possible concern regarding this design is that we may be measuring a general trend among the treated restaurants instead of a causal effect of the Michelin star changes. To alleviate this concern, we conduct a falsification test by generating a "placebo" guidebook publication date that is 90 days before the actual publication date. We then replicate Table 2.7 with the placebo guidebook date.[29] Results are presented in Column (2) of Table 2.22. The insignificant ATTs indicate that our results are unlikely driven by a general time trend.

[29]The re-construction of 18-month data with nine consecutive two-month blocks around "placebo" guidebook publication date results in 136 (out of 252) treated units, including 83 units for gaining Michelin stars and 53 units for losing Michelin stars, and 4,307 control units.

Table 2.21: Robustness Checks: SCM-DiD with Alternative Sentiment Measure and Alternative Window

	Alternative dependent variable (Section 2.6.2)	Alternative window (Section 2.6.3)
	(1)	(2)
After	–0.019**	–0.079***
	(0.008)	(0.029)
After × Increase	0.010	0.019
	(0.034)	(0.076)
After × Decrease	0.089***	0.039***
	(0.033)	(0.075)
Other control variables (Michelin stars, the number of reviews, cumulative average review rating, cumulative rating variance, Google search volume)	Yes	Yes
Pair–window FE	Yes	Yes
Restaurant FE	Yes	Yes
Observations	892	832
Number of pairs	223	208
R^2	0.821	0.821

Note: Robust standard errors clustered at pair level are in parentheses. *** p<0.01,** p<0.05.

Table 2.22: Robustness Checks: SynthDiD with Alternative Sentiment Measure and Falsification Test

	Alternative dependent variable (Section 2.6.2)	Falsification test (Section 2.6.4)
	(1)	(2)
Increase	–0.012	0.041
	(0.033)	(0.075)
Decrease	0.159***	0.052
	(0.062)	(0.133)
Total number of treated units	148	136
Total number of control units	4,334	4,307

Note: Overall ATT reported. Aggregated standard errors are in parentheses. *** p<0.01.

2.6.5 Replication with NYC Restaurants

To investigate whether the effects generalize to other countries, we conduct a replication study in the context of New York City (NYC). The detailed data construction process is described in Appendix 2.8.5. Using the mean consumer review rating as the dependent variable, we replicate the analysis of review sentiment in Equation (2.3.1). There are two key differences compared to our main analysis. First, instead of using Google search intensity as a proxy of restaurant demand Z_{it}, we use OpenTable reservation data from Farronato and Zervas (2022) to proxy the demand for NYC restaurants. As described in Section 2.5.2, Farronato and Zervas (2022)'s data contain information on whether each restaurant had been fully booked (sold out) between 18:30 and 19:30 on a daily basis. Thus, we measured the average demand for each restaurant within the 90-day window by calculating the percentage of fully-booked days during that period. Second, we replace the pair–window fixed effect $\alpha_{p(i)w(t)}$ with the window fixed effect $\alpha_{w(t)}$ because we do not match a restaurant with its control due to the limited sample.

The estimation results are presented in Table 2.23. Column (1) controls only for Michelin star levels and fixed effects, and Column (2) adds the full set of controls. Both columns reveal that the estimated coefficient for $After \times Increase$ is not statistically significant, indicating that gaining Michelin stars does not lead to changes in consumer review ratings. However, the estimated coefficient for $After \times Decrease$ is significantly positive, suggesting an increase in consumer review ratings for restaurants that lost Michelin stars. These results align with our main analysis. It is worth noting that the NYC data set has a smaller sample size (20 Michelin star increases and 8 Michelin star decreases), which impacts the level of statistical significance. Nevertheless, the overall trend and direction of the effects remain consistent.

Table 2.23: NYC Replication Results: Effects of Michelin Star Changes on Sentiment of Consumer Reviews

	DV: mean review rating	
	(1)	(2)
After	0.000	–0.022
	(0.040)	(0.047)
After × Increase	–0.146	–0.131
	(0.117)	(0.130)
After × Decrease	0.236**	0.210*
	(0.116)	(0.108)
One Star	–0.043	–0.062
	(0.069)	(0.068)
Two Star	0.086	–0.002
	(0.096)	(0.124)
ln(number of reviews+1)		0.026
		(0.038)
Cumulative average rating		0.629***
		(0.201)
Cumulative rating variance		0.105
		(0.214)
Percentage of fully booked days		0.124
		(0.130)
Window FE	Yes	Yes
Restaurant FE	Yes	Yes
Observations	252	252
Number of units	126	126
R^2	0.612	0.629

Note: Robust standard errors clustered at restaurant level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

2.7 Discussion and Conclusion

Expert opinion exerts tremendous influence on the consumer journey, but its effect on consumer experience is ambiguous as it can give rise to both expectation and reputation effects. Favorable expert opinions can enhance the reputation of a business, and potentially improve consumer experience by guiding consumer opinions, but they may also harm consumer experience by raising consumer expectations. Likewise, while unfavorable expert opinions may harm the reputation of a business,

they also have the potential to improve consumer experience by lowering consumer expectations. We investigate the tension between the expectation effect and the reputation effect as a result of expert opinion through the lens of consumer reviews in the context of the restaurant industry and Michelin stars.

We apply two synthetic-control-based methods to identify the effect of Michelin star changes on the sentiment and content of consumer reviews. We find consistently that decreases in Michelin stars improve consumer review ratings, suggesting that the expectation effect of expert opinions is stronger than the reputation effect. Analyses on review content further show that service and "value for money" appear to be the key drivers of customer satisfaction, and when a restaurant is removed from the Michelin Guide or loses stars, consumers tend to become less demanding on service, and focus less on value for money. As noted earlier, prior work has never documented the fact that a lowered expert rating can lead to a better consumer experience. We demonstrate that these results are unlikely to be driven by supply-side responses to Michelin awards or demand-side responses unrelated to the expectation and reputation effects, such as the changes in the mix of consumers visiting the restaurant.

Our results also reveal potential explanations for the "Michelin curse," i.e., the downside(s) of gaining Michelin stars. We offer substantive managerial insights for restaurant managers, the Michelin Guide, and other firms providing experience goods as a whole. For restaurants, the evidence presented in this paper suggests that losing a Michelin star can lower consumer expectations, which can potentially improve consumer review ratings. Therefore, losing Michelin stars is not necessarily bad news for restaurants. Our findings also bear implications on potential marketing strategies in response to Michelin stars. First, our analyses of the topics of consumer reviews indicate that consumers pay more attention to service than to food or menu, thus restaurants can strategically streamline menus (e.g., fewer unique dishes in more types of menu) to balance menu variety and service efficiency.

Second, because consumers are less concerned about value for money when a restaurant loses Michelin stars, these restaurants can potentially offer premium dishes with expensive ingredients (e.g., caviar, truffles, saffron, and wagyu) to increase revenue. However, this may not be an effective strategy for restaurants that gain Michelin stars. Third, "wine" is a word associated with the overall experience (Topic 5 in Table 2.8), and "sommelier" is a word associated with this topic when focusing only on unique words (Table 2.26 in Appendix 2.8.3). Therefore, restaurants may benefit from putting more effort in the wine list, and hiring professional sommeliers to recommend wines to complement customers' tastes and to pair with their menu choices. In addition to enhancing consumer experience, this can also directly enhance profitability because there is evidence that alcohol sales account for more than 80% of the profit for most fine-dining restaurants.[30] Finally, because consumers tend to focus more on service after Michelin star changes, restaurants should devote more resources on staff training in order to maintain a high standard of service.

For the Michelin Guide, given the controversy on "consistency" as a criterion and the lack of transparency in award decisions, the Michelin Guide can balance consistency and innovation (Ospina 2018) in their evaluation criteria. In addition, the Michelin Guide was established in the early 20th century and began to award stars for fine dining establishments in 1926. In the age of social media, consumer reviews and feedback can potentially be a valuable consideration in the assessment process.

For other businesses providing experience goods, our research offers valuable managerial insights. Companies tend to invest money and time with the purpose of being recommended by experts or showing better results in expert based rating systems. This often leads to businesses spending more on features and/or attributes

[30] https://www.thebalancesmb.com/restaurant-fine-dining-2888686

that are not necessarily relevant for the customer experience.[31] However, our findings reveal that winning such endorsements and/or awards does not always lead to improved consumer evaluations, and that losing an award may turn out to be a blessing in disguise. Essentially, businesses should be open to the understanding that favorable expert opinions can be a double-edged sword. As a result, they need to devote resources in a manner that balances "pleasing" experts (by playing to the criteria they use) and managing customer expectations and delivering fulfilling experiences.

There are several limitations to the present study that represent opportunities for future research. First, we focus on online consumer reviews and ignore other social media platforms and offline word-of-mouth. Incorporating other social media and offline word-of-mouth into the research framework would broaden our understanding of how consumer opinions are influenced by expert opinions. Second, due to the lack of access to sales and revenue data for UK and Ireland restaurants, we are unable to analyze the economic impact of (the change in) Michelin stars. Third, despite our efforts to understand the effect of Michelin star changes on the customer mix visiting a restaurant, our analyses are based on publicly posted reviews, which may not fully represent the actual customer base. Finally, this research mainly focuses on the Michelin Guide for Great Britain & Ireland with a replication study on New York City's Michelin Guide, and future research can extend the scope of the analyses to other countries and/or industries.

[31] See, for example, `https://www.forbes.com/sites/forbesbusinesscouncil/2022/03/30/how-hotel-art-affects-ratings/`

2.8 Appendix

2.8.1 Bootstrapped Standard Errors for SCM-DiD

We use the approach outlined by (Arkhangelsky et al. 2021) and (Adalja et al. 2023) to calculate bootstrap standard errors for the SCM-DiD analysis reported in Table 2.6. For each treated unit, we independently resample the donor pool consisting of control units 1,000 times. For each bootstrap sample b, the estimator $\hat{\delta}^b$ is obtained following the procedure described in Section 2.3.2. The bootstrap variance is calculated as $\hat{\mathcal{V}} = \frac{1}{1000}\sum_{b=1}^{1000}(\hat{\delta}^b - \frac{1}{1000}\sum_{b=1}^{1000}\hat{\delta}^b)^2$. The results shown in Table 2.24 are consistent with those in Table 2.6.

Table 2.24: Bootstrap Treatment Effect and Standard Errors

	Increase	Decrease
Estimation	0.086	0.363***
	(0.034)	(0.038)
p-value	0.227	0.000

Note: The table presents bootstrap mean, standard errors (in parentheses), and average p-value of the treatment effect among 1000 iterations. p<0.01, ** p<0.05, * p<0.1.

2.8.2 Additional Details on SynthDiD

Table 2.25 shows the cohort-level SynthDiD ATT estimates for review sentiment.

Table 2.25: Cohort-level Estimates by SynthDiD

Guidebook window	(1) Increase	(2) Decrease
2012	0.185	0.560***
	(0.122)	(0.332)
2013	–0.032	0.099
	(0.073)	(0.086)
2014	–0.048	0.236
	(0.084)	(0.186)
2015	–0.019	0.375
	(0.056)	(0.305)
2016	–0.047	0.169**
	(0.058)	(0.157)
2017	–0.012	0.192***
	(0.045)	(0.056)
2018	0.134*	0.557***
	(0.078)	(0.191)
2019	–0.021	0.336***
	(0.063)	(0.108)
2020	–0.094*	0.280***
	(0.049)	(0.104)

Note: Standard errors for each guidebook window calculated with bootstrap or placebo are in parentheses. We do not observe available treated units in guidebook window for the 2011 guidebook. *** p<0.01, ** p<0.05, * p<0.1.

2.8.3 Unique Words under the LDA Model

Table 2.26 displays the words that are unique to each of the five topics in decreasing order of the posterior probability.

Table 2.26: Unique Words under the LDA Model (K = 5)

Topic number	Topic name	Unique words
Topic 1	Value for Money	price, bite, expect, better, quality, little, expensive, quite, much, disappoint, small, high, overall, find, portion, though, dont, although
Topic 2	Issues with Order	ask, arrive, tea, wait, waiter, minute, leave, tell, give, seat, sit, didnt, afternoon, bill, offer, waitress, another, people, glass, hour, day, bring, show
Topic 3	Menu and Food	main, starter, dessert, cook, steak, beef, fish, cheese, sauce, chocolate, bread, lamb, start, chicken, side, duck, pudding, meat, tasty, cream, chip, roast, pork, scallop, follow, share, crab, salad, potato
Topic 4	Service and Staff	recommend, amazing, friendly, definitely, attentive, fantastic, highly, love, thank, beautiful, worth, perfect, night, return, cocktail, treat, always, welcome, superb
Topic 5	Overall Experience	tasting, star, every, dining, chef, win, michelin, present, year, kitchen, room, list, sommelier, ever, without, work, choice

2.8.4 Additional Details on the Reviewer-level Analyses

Reviewer Characteristics

Table 2.27 presents an example of reviewer characteristics at the time of the review.

Table 2.27: Reviewer Characteristics at the Time of the Review (Example)

Order of review	Local consumer	Picky consumer	Cum. # of restaurants until each awarded restaurant	Cum. mean review rating until each awarded restaurant
5	0	0	4	4
8	0	0	7	4.14

Restaurants Characteristics

We illustrate the process of computing restaurant characteristics at the time of the review. Table 2.28 shows an example of a reviewer with eight reviews. Columns (1) and (2) show the restaurant ID and review date. For each restaurant, we extract its time-invariant characteristics (i.e., price level and cuisine type) from the corresponding TripAdvisor page, as presented in Columns (3) and (4). Then, leveraging the dataset of 79 million reviews we have collected, we calculated the cumulative review characteristics for each restaurant up to the review date. For instance, for the first review in Table 2.28, we calculated the review characteristics for restaurant "105866" until 13 August 2014. Columns (5) to (7) illustrate these characteristics, including the logarithm of the total number of reviews, the mean and the standard deviation of previous ratings.

Assuming that the first five reviews in Table 2.28 are written by the same reviewer. For each review within her profile, we calculate the cumulative restaurant characteristics that have been reviewed up to that point. Table 2.29 provides an illustrative example. Columns (1) to (4) display the TripAdvisor profile of this re-

viewer, with each review indicating a review rating, the reviewed restaurant, and a specific date. Columns (5) to (11) describe the cumulative restaurant characteristics within the reviewer's timeline.

For instance, in the case of the first review in the profile, no cumulative restaurant characteristics exist. For the third review, the cumulative restaurant characteristics would incorporate information from the antecedent two restaurants. As detailed in Table 2.28, the third review corresponds to a "seafood" restaurant with "$$$$" price level. Before this entry, the reviewer had visited a "$$$$" priced restaurant and another priced at "$$ - $$$ ". Thus, by the third review, there are two unique price levels, shown in Column (5) of Table 2.29. In terms of cuisine type, uniqueness is determined based on specific word. For example, the first restaurant is labeled as "French, European," whereas the second is simply as "European" which is a subset of prior cuisine type. Therefore, up to the third review, the cumulative number of unique cuisine types is one (Columns (6) in Table 2.29). Moreover, Columns (7) to (9) in Table 2.29 compute the average review characteristics for those restaurant that have been reviewed so far. These calculations are derived from the information in Columns (5) to (7) of Table 2.28, respectively. We also determined the range of review ratings up to each respective review. Columns (10) and (11) in Table 2.29 present the rating range as "$mean \pm SD$" and "$mean \pm 2 * SD$", respectively, with both ranges derived using columns (8) and (9).

Table 2.28: Restaurant Characteristics at the Time of the Review (Example)

	Restaurant ID	Date	Price Level	Cuisine Type	ln(Cum. number of reviews+1)	Cum. mean rating	Cum. rating s.d.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
# 1	105866	2014-08-13	$$$$	French, European	6.864	3.889	1.056
# 2	033473	2014-08-25	$$ - $$$	European	5.357	3.923	0.946
# 3	086008	2014-12-03	$$$$	Seafood	3.996	4.450	0.828
# 4	018994	2015-02-11	$$$$	Bar, British	5.234	4.232	1.008
# 5	008075	2015-05-06	$$$$	Steakhouse	6.768	4.591	0.843
# 6	025005	2015-05-16	$$$$	European	7.001	4.535	0.827
# 7	037418	2015-08-15	$$$$	America	6.743	4.215	1.026
# 8	140968	2015-08-18	$$$$	Japanese	5.727	3.478	1.142

Table 2.29: Cumulative Characteristics of Restaurants Reviewed at the Reviewer Level (Example)

TripAdvisor Profile				reviewer-level cum. restaurant characteristics					
Order of Review	Review Rating	Restaurant ID	Date	Cum. number of unique price levels	Cum. number of unique cuisine types	Cum. average number of reviews	Cum. mean rating	Cum. rating standard deviation	Cum. rating range $mean \pm SD$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	4	105866	2014-08-13	–	–	–	–	–	–
2	4	033473	2014-08-25	1	1	6.863	3.889	1.056	[2.833, 4.945]
3	4	086008	2014-12-03	2	1	6.111	3.906	1.001	[2.905, 4.907]
4	4	018994	2015-02-11	2	2	5.405	4.087	0.943	[3.144, 5.000]
5	5	008075	2015-05-06	2	3	5.363	4.124	0.960	[3.164, 5.000]

Note: As review rating is in 5-point scale, the right boundary of the rating range is "minimum(5, $mean + SD$)" in Column (10).

2.8.5 Data Construction for Replication Study with NYC Restaurants

We construct the dataset for the replication study through five steps, outlined in Table 2.30. Steps 1-3 are conducted in the same manner as described in Section 2.5.2: focusing on the Michelin guidebooks in NYC from 2013 to 2017, matching the awarded restaurants with (Farronato and Zervas 2022)'s OpenTable reservation data, and retaining units that had reservation information available for both the pre-guidebook and post-guidebook windows. Moving on to the fourth step, we proceed to collect OpenTable reviews specifically related to these awarded restaurants. As a result, our NYC replication dataset consists of 73,229 reviews for 52 (out of 70) awarded restaurants from 2013 to 2017. The remaining 18 awarded restaurants are excluded from the dataset due to the absence of an OpenTable page. In the fifth step, we focus on restaurants where we observed reviews within both a 90-day window before and after the guidebook release date. In the end, the data set for NYC replication includes 126 "restaurant-guidebook year" units that correspond to 47 awarded restaurants. Among these 126 units, 20 experienced increases in Michelin stars, 8 experienced decreases in Michelin stars, while the remaining units maintained the same Michelin stars.

Table 2.30: Data Construction Steps in New York City Replication

Steps	Data sources	# Awarded Restaurants	# Michelin star increases	# Michelin star decreases	# "restaurant-guidebook year" unit
Step 1	Michelin Guides in NYC from 2013 to 2017	117	54	39	468
Step 2	OpenTable reservation data (Farronato and Zervas 2022)	70	28	13	230
Step 3	Reservations available on both sides of the guidebook release date	70	27	13	222
Step 4	OpenTable review data from 2013 to 2017	52	21	8	141
Step 5	Reviews available on both sides of the guidebook release date	47	20	8	126

Chapter 3

Expert Opinion on Supply-Side Responses

3.1 Introduction

In the preceding chapter, our focus was centered on understanding consumer response to expert opinions as reflected in customer reviews. This chapter, however, shifts focus to examine how restaurants have historically adjusted their menus in response to gaining/losing Michelin stars. In this chapter, we address the research question of how do restaurants respond to Michelin star awards, and what might have been improved given the effect of Michelin star status on consumer reviews? By answering these questions, we can provide managerial implications for restaurant managers.

The study closest to ours is Sands (2020), who examines the effect of receiving a Michelin star on restaurant survival, and finds that being awarded a Michelin star can increase the probability that a restaurant go out of business in subsequent periods. Our paper differs from Sands (2020)'s paper in three ways. First, we consider the annual changes in Michelin stars (i.e., both increase and decrease) from 2010 to 2020, while Sands (2020) only considers the star changes when the

restaurants receive stars for the first time. Second, we use time-varying menu prices from restaurants historical menus rather than average price levels from The New York Times. Third, we use synthetic control to test the robustness, which guarantees parallel pretreatment trends between awarded restaurants and control restaurants.

3.2 Data

In order to analyze restaurant responses to Michelin stars, we retrieved all available historical menus for each awarded restaurant since the publication of Michelin Guide 2010, using the website archival tool Wayback Machine (`https://archive.org/web/`) on the restaurants official website. We were able to obtain historical menus for 266 of the 271 Michelin restaurants. For the remaining five restaurants, we did not observe a current menu on their official websites or historical menus from their websites archive.

A restaurant can provide multiple types of menu choices (e.g., lunch tasting set, dinner tasting set, etc.) on a given day. We refer to a specific type of choice as a "submenu" for the focal restaurant and day, and denote the collection of all available submenus on that day as a "menu." In other words, a menu consists of all of a restaurants submenus for a specific day. In total, there were 4,156 menu changes and 12 types of submenus for the 266 Michelin restaurants from guidebook year 2010 to guidebook year 2020. The 12 types of submenus are: a-la-carte, quick lunch set, lunch tasting set, quick dinner set, dinner tasting set, seasonal, occasion (e.g., Valentines Day, Mothers Day, etc.), vegetarian, vegan, children/kids, pre-theatre, pescatarian. Among the 12 types of submenus, the quick lunch set and the dinner tasting set were the most popular options. The quick lunch set typically includes two or three courses, and the dinner tasting set provides a full dining experience with at least five courses.

We then manually organized the retrieved menus into a structured format. For

each menu collected, we extracted the earliest date that the menu was archived, the types of submenus it included, the number of dishes for the set menu, the name and ingredients for each dish, the price for each dish in the a-la-carte menu, and the price for the set menu. Because of the extensive amount of manual work required for organizing the menu data, we did not collect detailed menu items for the control restaurants but only checked the dates on their menu changes.

For each restaurant, we calculate the menu duration as the number of days between the last day the previous menu was archived and the earliest day when the focal menu was archived. For each menu, we count the number of submenu types and the number of submenu choices. For example, if a restaurant provided two quick lunch sets (£30 for two courses and £40 for three courses) and two dinner tasting sets (£100 for five courses and £120 for seven courses) on October 11, 2018, then under our definition, it provided two types of submenus and four submenu choices on that date. We choose the lowest price as the submenu price because it represents a "minimum spend" for that particular submenu. Thus, in the preceding example, the price for the quick lunch set is £30 and the price for the dinner tasting set is £100. We identify unique dishes within a menu based on the dish name and ingredients. Table 3.1 presents the summary statistics of the menu data. In general, one-star and two-star Michelin restaurants changed their menu more frequently, and restaurants with more Michelin stars tended to provide fewer unique dishes at higher prices, though the differences are not statistically significant due to the large standard errors.

3.3 Empirical Model and Results

This section examines the effect of Michelin star ratings on restaurant menus and discusses how a restaurant should adapt its marketing strategies in response to Michelin star awards. As described in the Section 3.2, we have collected 4,156

Table 3.1: Summary Statistics of the Menu Data

Guidebook year	# Michelin restaurants			
	no-star	one-star	two-star	three-star
Number of restaurants	128	231	29	5
Number of changes	823	2,967	281	84
Avg menu duration in days (s.d.)	203.40	161.94	160.15	232.83
	(211.62)	(126.81)	(143.28)	(262.69)
Avg number of submenu types (s.d.)	2.50	2.64	2.72	2.77
	(1.25)	(1.23)	(1.14)	(1.75)
Avg number of submenu choices (s.d.)	3.68	3.96	4.27	3.37
	(2.25)	(2.36)	(2.58)	(2.27)
Avg number of unique dishes (s.d.)	30.96	29.99	25.43	22.99
	(18.28)	(17.54)	(18.88)	(17.47)
Avg price for quick lunch set (s.d.)	29.07	32.49	51.46	58.59
	(10.12)	(12.13)	(19.32)	(10.68)
Avg price for dinner tasting set (s.d.)	68.57	74.66	129.86	190.74
	(21.65)	(21.50)	(25.61)	(72.25)

detailed menus with 12 types of submenus for 266 of the 271 awarded restaurants since the publication of Michelin Guide 2010. As there is evidence that diners may prefer variety to simplicity on restaurant menus,[1] we focus on variety and price in our analysis of the menus. For variety, we consider menu variety (the number of submenu types, and the number of submenu choices) and dish variety (the number of unique dishes). For price, we consider the prices of the quick lunch set and dinner tasting set, the two most popular submenu options. Further, we consider whether the restaurant provides seasonal submenus or occasion submenus, and whether the restaurant provides tailored submenus accommodating consumers' dietary requirements (e.g., vegetarian, vegan, children/kids, or pescatarian).

Because the Wayback Machine does not include all web pages ever published, and the frequency of snapshots that are captured varies across websites (in our case, menu snapshots are typically not captured on a daily basis), there are two types of

[1] https://www.forbes.com/sites/darrentristano/2015/02/03/restaurant-consumers-value-variety-over-simplicity-on-menus/

noise in our menu data: 1) the menus we collected are a subset of all of the menus that these restaurants ever had, and 2) the archived dates (and thus the start and end dates) for each menu may not be accurate. Therefore, when estimating the effect of Michelin star changes on menus, we no longer focus on a tight time window as in the prior analyses, but instead consider all menus within a guidebook year and estimate the following model:

$$\begin{aligned} Response_{ijt} = \quad & \beta_1 Increase_{jt} \times OneStar_{jt} + \beta_2 Decrease_{jt} \times OneStar_{jt} \\ & + \beta_3 Increase_{jt} \times TwoStar_{jt} + \beta_4 Decrease_{jt} \times TwoStar_{jt} \\ & + \beta_5 Increase_{jt} \times ThreeStar_{jt} + \alpha_j + \tau_t + \varepsilon_{ijt} \end{aligned} \tag{3.3.1}$$

where $Response_{ijt}$ is the outcome of interest (e.g., menu variety, dish variety, whether the restaurant provides seasonal or occasion submenus, whether the restaurant provides tailored submenus accommodating consumers' dietary requirements, and menu price) for menu i of restaurant j in guidebook year t. $Increase_{jt}$ and $Decrease_{jt}$ are indicator variables denoting the Michelin star change for restaurant j in guidebook year t compared to the previous guidebook year. $OneStar_{jt}$, $TwoStar_{jt}$ and $ThreeStar_{jt}$ denote the star level of restaurant j in guidebook year t. We allow the effect of the star change to vary by the current star level. α_j is a restaurant fixed effect for capturing unobservable restaurant characteristics, and τ_t is a guidebook year fixed effect for capturing unobservable time-varying factors. ε_{ijt} is an idiosyncratic error term.

Table 3.2, Columns (1) and (2) present the estimation results for menu variety, and Column (3) presents the estimation results for dish variety. Among the one-star restaurants, those with more stars in the previous year provided significantly more types of submenus and more submenu choices, but fewer unique dishes than those that had one star in the previous year. Those with no stars in the previous year provided more submenu choices, although the effect is marginally significant. In other

words, restaurants with a decreased star rating are more likely to streamline their menus (increased menu variety but decreased dish variety), possibly because such changes tend to improve efficiency in service and increase diners overall satisfaction.[2] This explains our estimation result for consumer reviews (Table 2.10 and Table 2.11 in Chapter 1) that restaurants with a decreased star rating saw an increase in their overall star rating and an increase in the proportion of Topic 4 (service and staff) in their review texts. Even if consumers preferred greater menu variety, they may not have noticed this decrease in variety because they became less concerned about menus and food ingredients (Topic 3).

Column (4) shows the estimation results for whether a restaurant provides seasonal or occasion submenus, and Column (5) shows whether the restaurant provides submenus tailored to specific dietary requirements. Compared with one-star restaurants that had the same star rating in the previous year, one-star restaurants that had either no stars or more stars in the previous year were more likely to provide menus specifically designed to accommodate consumers' dietary requirements. One possible reason is that by providing personalized menus catering to consumers' specific needs, restaurants are able to enhance their overall dining experiencebecause "menu" is the word with the highest probability for Topic 5 (overall experience) in Table 2.8.

Columns (6) and (7) of Table 3.2 show the estimation results for the prices of quick lunch sets and dinner tasting sets. Surprisingly, one-star restaurants with no stars in the previous year tended to lower their menu prices after gaining a star, while one-star restaurants with more stars in the previous year tended to increase their menu prices after losing stars. Among two-star restaurants, those with fewer stars in the previous year tended to lower their quick lunch set prices and increase their dinner tasting set prices. These menu price adjustments are possibly driven in part

[2] https://www.revenuemanage.com/en/insights/reduced-restaurant-menus-offer-opportunity-for-profit-and-simplicity/

by restaurants' intention to "neutralize" consumers expectation levels. However, the results in Table 2.10 and Table 2.11 in Chapter 1 show that the changes in Michelin star ratings are unlikely to influence the proportion of Topic 1 (Value for Money) in consumer review texts, suggesting that price may be less effective for enhancing consumer satisfaction after a change in the Michelin star rating.

There is evidence that Michelin-starred restaurants may include expert-selected ingredients that are expensive.[3] We further check whether the changes in Michelin star status incentivized restaurants to use expensive ingredients (e.g., caviar, truffles, saffron, and wagyu)[4] in their menus, but do not find a difference.

Together, our analyses of restaurant menus provide implications for restaurant responses to Michelin star rating changes. When the Michelin star rating of a restaurant changes, consumers tend to focus more on service than on the menu or menu items. Although a menu with a reduced number of dishes provides consumers with fewer alternatives, it can increase service efficiency and improve consumers' dining experience. This is possibly one of the reasons why awarded restaurants with a decreased star rating received better consumer reviews after the star decrease. However, restaurants with an increased star rating tended to focus more on price than on menu structure, leading to complaints about service in the consumer reviews. Our results suggest that in response to Michelin awards, a restaurant can streamline its menu structure to improve consumers' overall satisfaction, but price may not be an effective strategic variable for improving customer satisfaction.

[3] https://www.souschef.co.uk/blogs/the-bureau-of-taste/the-uk-s-most-michelin-ingredients

[4] https://www.lovefood.com/gallerylist/52001/the-worlds-most-expensive-ingredients-foods-2020

Table 3.2: Effects of Michelin Stars on Restaurant Menus

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Number of submenus types	Number of submenus choices	Number of unique dishes	Has seasonal or occasion submenus	Has dietary tailored submenus	Quick lunch set price	Dinner tasting set price
Increase × OneStar	0.085	0.296*	1.072	–0.022	0.110**	–1.006*	–3.492*
	(0.092)	(0.159)	(1.391)	(0.032)	(0.052)	(0.939)	(1.839)
Decrease × OneStar	1.859***	0.932**	–8.549**	–0.058**	0.716***	8.091***	6.580*
	(0.517)	(0.448)	(3.824)	(0.029)	(0.242)	(1.684)	(3.942)
Increase × TwoStar	–0.335	–0.512	–23.078	0.006	0.076	–2.419**	40.484**
	(0.335)	(0.333)	(14.087)	(0.077)	(0.087)	(1.129)	(20.255)
Restaurant FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Guidebook year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4,156	4,156	4,156	4,156	4,156	4,156	4,156
Number of restaurants	266	266	266	266	266	266	266
R^2	0.704	0.711	0.611	0.551	0.677	0.863	0.892

Note: Robust standard errors clustered at restaurant level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1. The variable "Decrease × TwoStar" is omitted because there are no such star changes during the data period. The variable "Increase × ThreeStar" is omitted because there is only one restaurant whose rating increased to three stars, and this restaurant did not have an archived menu from before this change.

3.4 Conclusion

In this chapter, we analyze restaurants historical menus to explore how the restaurants responded to Michelin star awards. We find that one of the reasons why restaurants with decreases in Michelin stars received higher star ratings after the star decrease is that they streamlined their menu structure and thereby improved the service quality. In contrast, restaurants with an increase in Michelin stars tended to focus on menu price rather than the menu structure, which led to complaints about service in consumer reviews.

Therefore, the following strategies can potentially benefit restaurants in response to Michelin awards: 1) strategically streamline menus to balance menu variety and service efficiency; 2) provide more personalized options catering to consumers specific needs (e.g., dietary, appetites), but with fewer different dishes; and 3) devote more resources to train staff on maintaining a high standard of service.

Overall, our results suggest that generally, in response to Michelin awards, an effective strategy for improving consumers' overall satisfaction is to streamline the menu structure, and that price changes may be less effective in this regard.

Although our study focuses on Michelin star restaurants, the implications extend far beyond this elite dining sector, touching upon universal themes in business and creative industries. First, our research highlights how Michelin star restaurants innovate in menu design and customer experience. This serves as a model for adaptability and innovation that is applicable across various sectors. Businesses, regardless of size or scope, can learn from these principles to enhance their own product offerings and adapt to changing market demands. Second, the pursuit of a Michelin star is, at its core, a pursuit of excellence and creativity. These are values that resonate across all industries. By studying the methods and impacts of such high standards, we provide a blueprint for organizations striving to differentiate themselves through creative excellence. Third, this study underscores the importance of customer experience, a crucial factor in the success of any service-oriented business. The insights gained can help other sectors understand how meticulous attention to detail and personalized service contribute to overall customer satisfaction and loyalty. Fourth, the presence and operations of Michelin star restaurants can significantly influence local economies, from tourism to local employment and the promotion of regional products. This aspect of our research may interest policymakers and economic developers in diverse regions.

Chapter 4

Spend Analysis 4.0: Automating Procurement Practices using Artificial Intelligence

4.1 Introduction

Procurement in large manufacturers is a complex operation, often involving the purchase of tens of thousands of products from thousands of suppliers, with annual purchase costs in the hundreds of millions of dollars. In 2020, US manufacturers alone spent $2.8 trillion in procurement, more than half of their revenue (U.S. Census Bureau 2022). However, such large-scale procurement processes often lack transparency across the entire company and rely on a medley of heterogeneous legacy systems that operate in silos, many of which were assimilated through previous mergers and acquisitions. The complexity of regularly updating the digital infrastructure and procedures of even a single business unit, let alone standardizing them across the entire corporation, is daunting. This situation results in many inefficiencies and missed opportunities. For example, it is typical to observe that the same types of products are procured from different suppliers at significantly different prices.

As a result, manufacturers conduct **spend analysis** to review their purchases and identify the greatest opportunities to save time and money by renegotiating supplier costs or redesigning products. As a procurement manager, the main reason for performing a spend analysis is to achieve your cost-reduction objectives. A spend analysis assists in meeting these goals by pinpointing the cost outliers and the most promising areas for cost reduction efforts. This process relies on procurement experts' know-how and is often performed manually. Thus, a spend analysis is typically performed manually (with the aid of spreadsheets) by procurement experts, who possess the requisite nuanced industry knowledge to understand the nature of procurement transactions and read between the lines. As such, this process is both time-consuming and expensive, and often involves multiple external procurement experts and can take several months to complete. Given the impracticality of manually filtering through thousands of suppliers, the scope of such manual analysis is often limited to a subset of suppliers with high procurement spending. This approach can be biased and prone to errors, ultimately limiting the identification of cost-saving opportunities.

A natural evolution in enhancing procurement practices, consistent with the theme of Industry 4.0, would be to automate (or semi-automate) spend analysis (Olsen and Tomlin 2020). By adopting an automated approach, businesses can easily integrate, connect, and interface their procurement practices. This can be particularly advantageous for large manufacturers frequently acquiring or restructuring new business units, or SMEs (Small and Medium-sized Enterprises) facing cost barriers for spend analysis. As such, this proposition has generated substantial attention during the pandemic in general (Dittrich et al. 2020), and significant investments from leading consulting firms in particular (McKinsey 2021, Garcia 2021).

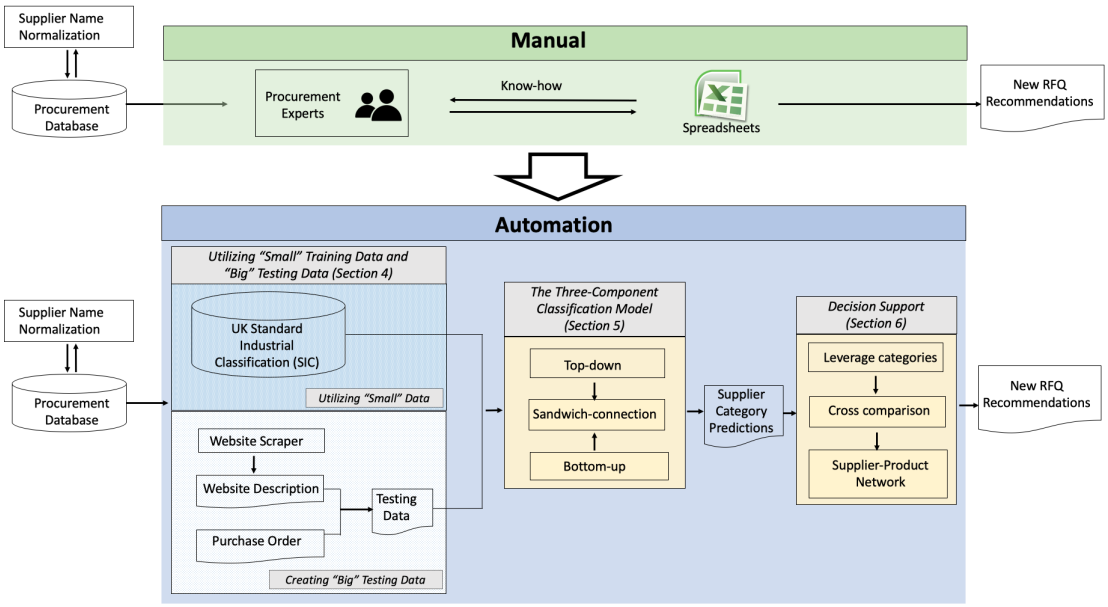
However, while machine learning and natural language processing (NLP) are well-suited for automating these tasks, integrating these technologies into procurement to match the nuanced insights of experts presents four significant challenges.

First, there exist no true hierarchical category labels for any given set of suppliers. In other words, no third-party organization accurately classifies unlisted suppliers into specific industry categories, and existing self-reported categories (e.g., in the case of the UK) are highly inaccurate and unreliable for practical use. Second, for any given supplier category, sufficiently large pre-existing training data does not exist. For example, for any group of suppliers providing specific product categories, there may be at most a few dozen suppliers, which is not enough to train a reliable classifier. Third, the automatic spend analysis needs to be flexible and applicable in different contexts, not just in a specific supplier category. Since machine learning classifiers depend heavily on the training data set, a hierarchical classifier developed for one manufacturer may not be appropriate or applicable to another. Finally, building a hierarchical classification algorithm that is accurate beyond two levels of hierarchy is fundamentally difficult because classification errors propagate between levels (Dumais and Chen 2000). To identify savings opportunities accurately in practice, however, five to six levels of hierarchical sub-categories are needed.

In this paper, we present a comprehensive methodology that employs NLP and machine learning to automate spend analysis that successfully replicates the procurement expert's know-how. Our methodology categorizes suppliers' data into a detailed and extensive hierarchical taxonomy with 6 levels and 15,574 distinct categories. This structured categorization helps to identify leverage suppliers that can result in realized cost savings of 5-10% of the current invoice values when request-for-quote (RFQ) process is initiated. The architecture is depicted in Figure 4.1, and we provide a detailed expansion for each main section in the subsequent paragraphs.

Our methodology begins with addressing the lack of proper training data (Section 4.4). To do so, we utilize "small data" from a detailed hierarchical taxonomy structure and "big data" to enrich the testing data. Namely, we train a classifier to learn the UK's Standard Industrial Classification (SIC) guide (UK Office for National Statistics 2007), a standardized hierarchical taxonomy guide describing

Figure 4.1: Detailed Architecture of the Spend Analysis Automation System



the company's nature of business. It can be easily replaced by other standardized taxonomy guides such as the United Nations Standard Product and Services Code (UNSPSC), the North American Industry Classification System (NAICS), or any company's internal taxonomy. This "taxonomy" can be an evolving document that manufacturers can customize over time, e.g., add more detailed descriptions for their own procurement needs. This "small data" from hierarchical taxonomy mirrors the procurement expert's logic and intuition that are applied when classifying suppliers. Then, we enrich the suppliers' information by incorporating web-scraped texts (containing general contextual supplier information and detailed product information) in addition to the raw purchase order data from the procurement database. This "big data" mirrors procurement experts' general contextual knowledge and familiarity with specific product descriptions.

Next, in Section 4.5, we propose a three-component classification model. It utilizes (a) traditional top-down classification, effective for higher levels of the taxonomy (e.g., level 1); (b) traditional bottom-up classification based on word similarity, targeting lower taxonomy levels (e.g., level 6); and (c) an innovative "sandwich-

connection" component that merges predictions from the two traditional methods, leveraging parent-child node relationships within the taxonomy. Our three-component classification model identifies the most likely categories for each supplier through levels 1 to 6. This creates a unique "DNA" profile for each supplier, represented as a binary vector that includes all sub-categories from levels 1 to 6. To evaluate our model's accuracy, we derived a partial list of true labels for a small, randomized subset of suppliers through a series of carefully designed experiments, utilizing both crowdsourcing and expert validation. Against these best estimates of the true labels,we demonstrate that the three-component model significantly enhances prediction accuracy (as measured by the F1 score) compared to existing benchmark models in hierarchical classification.

Lastly, in Section 4.6, our methodology is accompanied by a decision support tool that can convert the classification into potential savings opportunities. It performs a Kraljic analysis (Kraljic 1983) to identify the "leverage" categories of suppliers (those with low risk and high economic volume). Moreover, using the supplier's DNA, it enables the cross-comparison of all suppliers to identify a supplier's competitors or find a list of suppliers that can provide certain products. This approach provides a detailed understanding of the supplier-product networks within manufacturers, aiding in the strategic selection of suppliers. It also helps in crafting targeted requests-for-quote (RFQs), which can lead to potential cost savings through volume or price discounts.

We report the implementation of our automated spend analysis in Cranswick plc, a leading food manufacturer in the UK. It provided us with detailed data on its procurement transactions over two years from January 2019 to December 2020, which amounted to a total invoice value of £1.571 billion. Our automated spend analysis was able to examine all 2,170 suppliers and accurately classify them into hierarchical categories. Together with the decision support tool, our methodology was instrumental in identifying the "leverage supplier categories and generating a

list of target suppliers to issue RFQs and the estimated cost-savings within *days* (instead of months).

If Cranswick plc follows through on the RFQ recommendations, significant cost savings are achievable. To estimate the cost-savings that are attributable to automation, we performed a simulation analysis (based on a model calibrated using Cranswick plc data) that incorporates (i) an improved scope of analysis, (ii) increased accuracy in classifications, and (iii) increased frequency of spend analysis performed. Specifically, over a two-year period, the automation can generate *additional* savings of 2-3% of total procurement costs compared to traditional manual spend analysis methods. This translates into £16-22 million in annual savings, underscoring the significant financial advantages of adopting automated spend analysis techniques.

The contribution of our paper are two-fold. First, our paper introduces a methodology that relaxes the reliance on extensive datasets commonly needed for machine learning algorithms, and instead train small but informative data efficiently. This approach enhances the flexibility of our methodology in manufacturer settings. For example, our methodology has been applied in a merger and acquisition setting where a German industrial manufacturer acquired a Swedish company. The German acquirer had a detailed internal supplier-classification database and taxonomy and wanted to classify the Swedish firm's extensive supplier list according to their existing system. The German company shared with us their own hierarchical taxonomy and their supplier database. Instead of using SIC taxonomy as the training data (as shown in "Utilizing 'Small' Data" in Figure 4.1), we utilized the German company's hierarchical database to train our three-component model. Our three-component model was then able to provide the classification of the Swedish firm's suppliers according to the German company's taxonomy, facilitating its timely integration.

Second, our methodology takes large sets of unstructured data and converts them

into a structured format that can provide strategic insights. This feature allows for applicability across various industries beyond manufacturing, particularly where structured data is limited but abundant descriptive documentation is available. For example, in the financial or legal services sectors, there are extensive documentations of regulation, codes of practice, and compliance reports. Our three-component model may map these processes into a hierarchical structure. Once the hierarchical structure of the regulations, sections, clauses, and sub-clauses are understood, the three-component model could be trained on the extensive regulatory documentation. It could then classify new documents (e.g., live cases) into the relevant categories within the regulatory framework. After categorization, the model pinpoints which sub-clauses or sections are most frequently associated with live cases. Such classification supports decision-making by highlighting areas that require attention, allowing for proactive measures to address compliance issues or service needs.

4.2 Literature Review

Our paper contributes to the field of hierarchical classification, specifically in the context of text data. This area of research delves into three types of classifier (Silla and Freitas 2011). The first approach, known as the top-down approach, involves constructing a separate flat classifier for each level of the hierarchy. This approach only predicts a node if its ancestor nodes have also been predicted, a strategy documented in various studies (Dumais and Chen 2000, Cesa-Bianchi et al. 2006, Esuli et al. 2008, Cerri et al. 2014). However, a notable drawback of the top-down approach is the propagation of classification errors from upper to lower levels within the hierarchy. This flaw becomes particularly problematic in hierarchies that extend beyond two levels, where errors at higher levels inevitably affect the accuracy of lower-level classifications.

The second approach, known as the bottom-up approach, starts by predicting

labels at the leaf nodes and then infers their ancestors' labels through heuristics (Ceci and Malerba 2007). However, this approach ignores the information about the parent-child relationship along the hierarchy, and thus leads to low accuracy when the number of leaf classes becomes large.

The third approach is the so-called big-bang approach where a single and comprehensive classifier is built to address the entire classification problem. Unlike top-down or bottom-up approaches that break down the classification problem into smaller tasks, the big-bang approach aims to leverage the full complexity of the data structure and class hierarchy from the outset. Most of the studies focus on optimizing traditional machine-learning algorithms (e.g., neural network, support vector machine, decision tree) and require a large set of pre-labeled training data (e.g., McCallum et al. 1998, Cai and Hofmann 2004, Peng et al. 2018, Mao et al. 2019), and is therefore infeasible in context when such training data is not available, as is the case in procurement.

Addressing these challenges, our innovative three-component classification model synthesizes the strengths of both the top-down and bottom-up approaches by using the "sandwich connection." By doing so, it achieves significantly improved accuracy across hierarchical structures that are both *deep* and *broad*. As a result, we introduce a classification methodology that is appealing to a wide range of practitioners. Our approach effectively overcomes the constraints associated with traditional classification methods, paving the way for precise and efficient classification of hierarchical data. To our knowledge, this is the first instance of creating a hierarchical classification model that operates without relying on pre-defined true labels to categorize suppliers on a large scale.

Accurate hierarchical classification is not merely an academic concern but a practical necessity in procurement, where selecting the right suppliers and products is critical for issuing RFQs effectively. The design of effective procurement mechanisms has long been a prominent area of research in the operations management field

(Vickrey 1961, Laffont and Tirole 1993, Elmaghraby 2000, Hasenbein et al. 2010). Recent developments have focused on identifying the optimal sourcing strategies under different market attributes and environments (e.g., Chaturvedi et al. 2014, Li and Wan 2017, Beil et al. 2018), as well as examining the best way to issue the RFQs (e.g., Beil and Wein 2003, Wan and Beil 2009, Duenyas et al. 2013). Our paper complements these theoretical studies by examining the *upper stream* concerns of *which* supplier/product categories among its vast suppliers to issue RFQs in the first place. Although the field of data analytics is burgeoning (Mišić and Perakis 2020), its reach has been somewhat limited in the procurement space. We contribute by incorporating industrial-level data to address practical procurement problems and bringing analytics into this classic operations space.

In the realm of raw transaction data, online shopping platforms such as eBay.com and Amazon.com have seen a dramatic rise in popularity, with millions of new items being added daily (Cevahir and Murakami 2016). To effectively manage this immense flow of invoice and procurement transactions, these companies typically organize the items into distinct categories (Shen et al. 2012). Such categorization is vital not only for enhancing user experience by streamlining search and navigation, but also for boosting operational efficiency, which aligns with the objectives of our paper. However, existing research primarily focuses on developing deep-learning models to analyze features based on large-scale procurement transaction data (e.g., Tarawneh et al. 2019, Akanksh et al. 2023). This approach does not align with our context, and we will outline the specific challenges in Section 4.3.2.

4.3 Problem Description

A spend analysis carried out by a manufacturer aims to create transparency in its procurement practice (i.e., the products purchased, the suppliers it purchased from, and the total quantity and cost of the purchase), identify opportunities for savings,

which could then be utilized to initiate RFQ processes to reduce costs (e.g., through private negotiations with suppliers or public auctions). To do so, one must gather and organize the procurement data, group the suppliers and products into (hierarchical) categories based on their similarities, and provide nuanced insights into the procurement process. In this section, we describe the challenges associated with conducting a spend analysis using Cranswick plc, a leading UK food manufacturer, as an example.

4.3.1 Cranswick plc's Transaction Data

Cranswick plc produces a range of fresh foods with a fully integrated supply chain. It sources from farmers, processes raw products, packages and labels products, and ships them globally. It is a member of the FTSE 250, with a total reported revenue of £2 billion (approx. $2.5 billion) in 2022 (Cranswick 2023).

Cranswick plc provided us with its procurement transaction data that they assembled from multiple internal data sources from 2019 to 2020 as a spreadsheet. Each row represents a purchase order, including product-related information (e.g., item description, item code, item price), order-related information (e.g., order date, order quantity, order currency, total invoice value, delivery date), buyer-related information (e.g., a business unit within Cranswick plc), and supplier-related information (e.g., supplier name). The suppliers are not publicly listed firms and do not have an official classification that is consistently referenced.

The data covers 556,866 procurement transactions with a total invoice value of £1,571 million across two years. There are 136,190 products (identified by item description) and 2,999 suppliers (identified by supplier name). Table 4.1 summarizes the key information of the raw procurement data by years.

To interpret the raw data effectively, it is necessary to standardize supplier names due to the frequent inaccuracies or variations in recording these names across dif-

Table 4.1: Summary Statistics of the Raw Procurement Data.

	2019	2020	Total
Number of suppliers	2,161	2,922	2,999
Number of products	68,998	78,277	136,190
Number of purchase orders	265,662	291,204	556,866
Number of business units (buyer)	12	13	13
Total invoice value (£million)	717	854	1,571

ferent business units (e.g., “ABC Limited," “Advanced Business Corp," “ABC"). Procurement experts achieve this by conducting a Google search for each listed supplier name and collecting the top website URL from the search results. When the search for two or more listed suppliers leads to the same URL, they are considered to be the same entity and are consolidated under the official name listed on the website. This verification process yielded 2,170 (from 2,999) unique suppliers, of which 1,921 had an official website URL.

4.3.2 Manual Spend Analysis: Relying on the Know-How of Procurement Experts

Figure 4.2: Screenshot of part of raw procurement data

A	B	C	D	E	F	G	H	I
Business_Unit	Order_Value	Order_Number	Item_Description	Item_Number	Order_Date	Supplier_Name_Client	Quantity_Ordered	Invoiced_Value
S18	40.88	PPO029411	Series 1600	N-MCO-0001	07/09/2018	ABC Limited	8	40.88
S18	51.66	PPO030171	Job 14874 end cap	N-MCO-0001	17/10/2018	ABC Limited	14	51.66
S18	20	PPO030171	Carriage	N-CAR-0001	17/10/2018	ABC Limited	1	20
S18	932.7	PPO030896	10 off raised profile as per drawing	C-P&M-0001	22/11/2018	ABC	10	932.7
S18	1170.76	PPO031546	Gearbox: g500-B110 Motor: MD 071-32	N-MCO-0003	02/01/2019	ABC	2	1170.76
S10	88493.6	1047627	JBS BRAZILIAN CBEEF 6.35KG	100858	21/12/2018	JBS GLOBAL UK LTD	3483	88468.2
S10	79756	1047759	JBS BRAZILIAN CBEEF 6.35KG	100858	02/01/2019	JBS GLOBAL UK LTD	3139	79730.6
S10	88493.6	1048103	JBS BRAZILIAN CBEEF 6.35KG	100858	15/01/2019	JBS GLOBAL UK LTD	3483	88468.2
S12	5.04		2 SIS ANG M&S HW BR 230-330g	33610	28/12/2018	2 Sisters Poultry Ltd	3219	16738.8
S12	5.04		2 SIS ANG M&S HW BR 330-380g	33610	28/12/2018	2 Sisters Poultry Ltd	1187.5	6175
S12	5.04		2 SISTERS ANGLESEY HW BREAST	33610	28/12/2018	2 Sisters Poultry Ltd	1850.5	9622.6
S12	5.04		2 SISTERS ANGLESEY HW BREAST 280-33	33610	28/12/2018	2 Sisters Poultry Ltd	2365	12281.78

In the process of manual spend analysis, procurement experts are tasked with understanding the scope of suppliers and their products from the raw procurement data. Figure 4.2 displays a screenshot of the raw procurement data in a spreadsheet, as received from Cranswick plc.

Understanding such data is deeply dependent on the nuanced industry knowledge of the experts, challenged by several factors. Firstly, a significant portion of the purchased products (Column D in Figure 4.2) are listed in a non-descriptive manner (e.g., "2 SIS ANG M&S HW BR," "JBS Brazilian CBEEF"), making identification difficult. Additionally, product descriptions can sometimes be misleading. For instance, an item listed as a "1200mm wide pretzel" might initially seem related to food items. Yet, an experienced procurement expert, recognizing the unusual size (1.2 meters), would correctly deduce that it refers to a part for a conveyor belt machine.

Secondly, raw procurement data often lack vital contextual details necessary for accurate supplier classification. For example, when a purchase order lists "boilers," it is challenging to know whether the supplier is a manufacturer, wholesaler, or service provider of boilers. Distinguishing the supplier's role is vital for a comprehensive understanding of procurement practices and supplier networks. Consequently, spend analysis extends far beyond merely identifying familiar terms. Adding to the complexity, suppliers may offer a diverse range of products and services. While one supplier might focus on a specific product, another could provide a broad array of goods and services. Recognizing the extent of a supplier's product range and service offerings is also important for negotiating price and volume discounts.

Thirdly, and perhaps most importantly, a procurement transaction indicates which product *was* purchased from a supplier, but it does not reveal which other products *could* have been purchased from it. Thus, two similar suppliers could appear very different based on the procurement data alone. As a result, drawing insights into supply networks from the procurement data requires procurement experts' familiarity with the industry, and their ability to read between the lines.

After fully analyzing the transaction data, it is necessary to classify suppliers into hierarchical groups according to their similarities. The objective is to determine which categories of suppliers to approach for an RFQ process, aiming to negotiate

savings via volume or price discounts. The deeper the hierarchy and the more granular categories, the better for gaining detailed insights for target identifications. Utilizing a hierarchical structure proves beneficial in practical scenarios, particularly because a manufacturers procurement operations are often decentralized by product categories. This approach facilitates the alignment of specific procurement teams with the corresponding categories for the execution of RFQs.

However, manually comparing thousands of suppliers, even with the help of spreadsheets, is difficult. Typically adhering to the Pareto principle, a procurement expert might concentrate on a limited subset of suppliers (e.g., 20%) that account for a significant portion of procurement expenses (approximately 80%). In the case of Cranswick plc, we find that 68 of 2,170 suppliers (3.1%) are responsible for 80% of the total invoice value. Consequently, a manual spend analysis would prioritize these 68 suppliers, constructing a detailed hierarchical classification centred around them, based on the procurement experts' knowledge. This process, driven by speculative hypotheses about the supply network, is inherently susceptible to biases and inaccuracies, overlooking the majority of suppliers and potentially missing out on savings opportunities. Moreover, the insights from one manufacturer are not transferrable to another, which requires the procurement experts to start from scratch when examining a new manufacturer. As such, this process is both time-intensive and costly, and is typically reserved only for large manufacturers with sufficient procurement spend volume to justify the costs.

4.3.3 Automation Challenges

Designing a "smart" methodology that can infer a procurement expert's nuanced understanding of the procurement setting from the vast procurement data (e.g., what specific items a particular supplier *could* produce) seems to be a natural evolution towards Industry 4.0. Although it may seem apparent that NLP and machine learning techniques could be employed, creating a hierarchical classification model in a procurement context faces four methodological challenges.

First, in a procurement environment, there is an absence of properly structured training data with predefined categories, meaning there are no existing *true* hierarchical category labels for each supplier to utilize. Second, developing a classifier that can accurately group a given supplier requires a significant volume of training data, often hundreds or thousands of samples. Yet, for each specific sub-category, the market might only offer a few dozen trustworthy suppliers, leading to a scarcity of *large* and *pre-labeled* datasets necessary for traditional machine learning approaches in procurement. Third, the effectiveness of machine learning classifiers is highly dependent on the dataset they were trained on, making a classifier designed for one sector (such as agriculture) potentially ineffective in another (such as heavy manufacturing). Fourth, even when there is enough pre-labeled data available, constructing an accurate multi-level hierarchical classification model that goes beyond two levels presents its own set of challenges. However, for the model to be of practical use, it needs to correctly classify suppliers across five to six hierarchical levels.

In the ensuing sections, we present a methodology that overcomes all of the above challenges and replicates the know-how of procurement experts.

4.4 Utilizing “Small” Training Data and “Big” Testing Data

Our aim is to find the most likely set of categories that a supplier belongs to. To address the challenge presented by the lack of extensive pre-labeled training datasets, we propose to train a classifier on a comprehensive 6-level Standard Industrial Classification (SIC) taxonomy, which we term as “small” training data. Then, we enrich the suppliers information from raw procurement data to create “big” testing data.

4.4.1 Utilizing “Small” Data

Expanded SIC with 6 levels.

We aim to develop a machine learning model that learns the details of the SIC guide. The SIC is a well-established taxonomy introduced by the UK government to classify business establishments by the type of their economic activity. It is a hierarchical supplier classification system, describing the activities and sub-activities in a hierarchical tree structure (a directed graph with each node having at most one parent node). The nodes in the tree contain category labels corresponding to the level of specificity. For instance, the highest level (i.e., level 1) categories represent the most general business activities, such as “manufacturing,” “agriculture,” “wholesale,” etc. Then, each level-1 category is broken down into more specific sub-activities at level 2. For example, the level-1 category “manufacturing” is broken into level-2 categories such as “Manufacture of food products,” “Manufacture of beverages,” and so on. Further, every level-2 category is broken down into level 3, and then into level 4. In some cases, the level-4 category is further broken down into level 5. Therefore, the lowest-level SIC categories can be either at level 4 or level 5 of the hierarchy, describing the most specific business activities, e.g., “butter and

cheese production,”“Growing of citrus fruits”, etc. For each lowest-level supplier category (i.e., level 4 or 5) in the guide, the SIC further details the types of *specific products* that the suppliers should carry.[1] For example, level-5 category “Butter and cheese production” contains eight detailed product categories: “Butter blending,” “Butter milk,” “Butter oil,”“Butter production,”“Butterfat,”“Cheese,”“Curd production,”and “Dairy preparation of cheese and butter.” Table 4.2 shows two examples of SIC hierarchical categories.

Table 4.2: Examples of SIC Hierarchical Categories.

	Example 1	Example 2
Level 1	Manufacturing	Agriculture
Level 2	Manufacture of food products	Crop and animal production, hunting and related service activities
Level 3	Manufacture of dairy products	Growing of perennial crops
Level 4	Operation of dairies and cheese making	Growing of citrus fruits
Level 5	Butter and cheese production	—
Level 6	Butter oil	Lemon growing

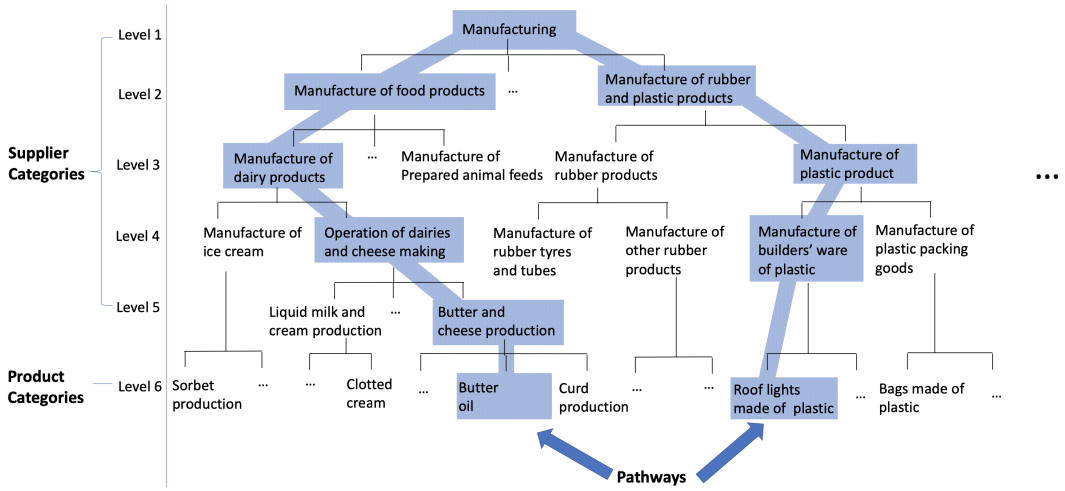
We define a SIC *pathway* as a sequence of nodes from the level-1 category to the level-6 category (see Figure 4.3). In the cases of original categories that end at level 5 (e.g., Example 1 in Table 4.2), the pathways in expanded SIC contain nodes in six levels; in the cases of original categories that end at level 4 (e.g., Example 2 in Table 4.2), the pathways in expanded SIC contain five hierarchical nodes without level 5.

In our three-component classification model, we design the hierarchical taxonomy to reflect the industry structures and classification logic used by procurement experts. While our approach utilizes the SIC system for demonstration, it is flexible and can be applied to other comprehensive taxonomies such as the United Nations Standard Products and Services Code (UNSPSC) or the North American Industry Classification System (NAICS), as well as to pre-existing internal supplier category databases. The chosen taxonomy, while serving as an initial framework, is intended

[1]The list of specific products can also be checked via `https://www.siccode.co.uk/`.

to be adaptable and evolve to meet the specific needs of manufacturers. Throughout the paper, we train the model to learn the expanded 6-level SIC taxonomy so that it can classify a supplier into appropriate pathways.

Figure 4.3: Expanded SIC Taxonomy with 6-level Categories.



SIC summary statistics.

Each node of the SIC tree is a category, and associated with it is a textual label. We construct the training data for each label by joining the texts of the category labels that appear in itself and all its child nodes. Take the left-hand-side pathway in Figure 4.3 as an example. The training data for its level 1 category "manufacturing" is the joint text from itself and all its child nodes in levels 2-6 (i.e., all the text labels in all pathways), and the training data for its level 6 category is the label itself (e.g., *"Butter oil"*).

To facilitate text data, we preprocess all training categories and training data by splitting the text into its component words, eliminating punctuation and numbers, lemmatizing words into dictionary form, and removing single-character words and stop words. Table 4.3 summarizes the statistics of SIC categories/labels and training data by hierarchical levels. SIC includes 21 level-1 categories and 15,574 level-6 categories, indicating 15,574 unique pathways. We find that 2,975 (out of 15,574)

pathways traverse all six hierarchical levels, whereas the remaining 12,599 pathways are without level 5. The average length of the training data is the number of words after preprocessing. We observe that higher-level categories tend to have longer training data than lower-level categories, which is intuitive due to the construction process above. Moreover, pathways that traverse all six levels tend to have longer training data than pathways that are without level 5.

Table 4.3: Summary Statistics of SIC.

	Number of categories/labels			Average length of the training data		
	Total	pathway traverse all six levels	pathway without level 5	Total	pathway traverse all six levels	pathway without level 5
Level 1	21	15	21	18.80	24.01	17.56
Level 2	88	43	87	16.56	21.69	15.35
Level 3	271	65	242	12.92	18.01	11.72
Level 4	614	78	536	8.99	13.86	7.83
Level 5	191	191	–	9.81	9.81	–
Level 6	15,574	2,975	12,599	3.95	3.86	3.97

Representing each category with a feature vector.

In the SIC taxonomy, let $\mathcal{C}_i$ denote the set of all categories in the hierarchy associated with level $i \in \{1, 2, 3, 4, 5, 6\}$, and let $|\mathcal{C}_i|$ denote the number of categories in each hierarchy i. For example, $\mathcal{C}_1$ represents the set of all $|\mathcal{C}_1| = 21$ categories at level 1, and $\mathcal{C}_6$ represents the set of all $|\mathcal{C}_6| = 15,574$ categories at level 6. We let $c_{ij} \in \mathcal{C}_i$ denote the j^{th} category at level i in the hierarchy. For example, an element $c_{1j} \in \mathcal{C}_1$ may represent a general business category (e.g., "manufacturing") at level 1; and an element $c_{6j} \in \mathcal{C}_6$ may represent a specific product category (e.g., "Butter oil") at level 6.

For any category in the SIC taxonomy $c_{ij} \in \mathcal{C}_i$ for $i = \{1, 6\}$, we create a corresponding feature vector.[2] For level-1 categories ($i = 1$), we first extract the

[2]One can similarly create feature vectors x_{ij} for each category $c_{ij} \in \mathcal{C}_i$, for $i = \{2, 3, 4, 5\}$, which can be utilized for evaluating the traditional top-down classification as a benchmark.

unique unigrams and bigrams from each of the 15,574 level-1 training data, which results in 38,189 features (i.e., 6,151 unigrams and 32,038 bigrams). Then, for each level-1 category $c_{1j} \in \mathcal{C}_1$, we compute the TF-IDF (term frequency-inverse document frequency) scores of each feature. To enhance machine learning model performance, we select those in the top 20^{th} percentile as the most informative features (Ramos et al. 2003, Domingos 2012), corresponding to 7,729 features (i.e., 2,342 unigrams and 5,387 bigrams). For each of the 21 level-1 categories $c_{1j} \in \mathcal{C}_1$, its *feature vector* $x_{1j} \in \mathbb{R}^{7729}$ is created by computing the mean TF-IDF scores for each feature from its training data.

For level-6 categories ($i = 6$), we also extracted all unique unigrams (but not bigrams) from 15,574 level-6 training data to emphasize product word similarity matching. It resulted in 6,123 unigrams. Again, for each of category $c_{6j} \in \mathcal{C}_6$, its corresponding feature vector $x_{6j} \in \mathbb{R}^{6123}$ is created from the TF-IDF scores for each feature from its training data.

4.4.2 Creating "Big" Testing Data

The testing data includes 2,170 suppliers from Cranswick plc's procurement transactions, and each supplier needs to be classified into a set of SIC hierarchical categories. As described in §4.3.2, the raw procurement data are often not very descriptive (e.g., "JBS CBEEF"), may even be misleading (e.g., "1200mm wide pretzel"), do not convey contexts of product and suppliers such as whether a supplier is a manufacturer, a wholesaler, or a service provider related to the product (e.g., "boilers"), and only indicates which product *was* purchased without revealing which products *could* be purchased.

To reflect the procurement expert's contextual understanding and familiarity with specific product descriptions, we enrich the suppliers' information with "big" data available on the web. Specifically, for each of the 1,921 suppliers with identified

URLs in the verification process of §4.3.1, we scraped its general business description and detailed product information from the official website. The added website data provides us with general descriptions of the suppliers, the contexts into products, and specific product descriptions that suppliers *could* provide but have not been recorded in raw procurement transactions.

A supplier m in Cranswick plc testing data is represented by its testing document d_m. For the 1,921 suppliers with official websites, we represent supplier m's testing document by $d_m \equiv (d_m^{gen},\ d_m^{spe},\ d_m^{PO})$, comprising of text data associated with the supplier's general business description obtained from the web (d_m^{gen}), its specific product description from the web (d_m^{spe}), and the purchased item descriptions from raw procurement purchase order data (d_m^{PO}). For the remaining 249 suppliers without official websites, $d_m = d_m^{PO}$.

Table 4.4 provides the summary statistics of the preprocessed testing data. For purchased order data in the raw procurement database, although the purchased item description is typically a long text with 474.78 words on average, it only contains 38.98 unique words. These long texts often arise from standardized instructions for restocking previous purchase orders. The limited text data from purchase orders have been significantly enriched by the scraped website data, which contains the general business description with an average of 68.15 unique words per supplier as well as the detailed product description with 90.98 unique words. Note that the testing data quality can be further improved by adding public LinkedIn page information or other private information from third-party organizations.

Table 4.4: Summary Statistics of the Enriched Testing Data.

	Purchased Item Descriptions (d_m^{PO})	General Business Description (d_m^{gen})	Detailed Product Description (d_m^{spe})	Suppliers' Testing Document (d_m)
Number of suppliers (after verification)	2,170	1,921	1,921	2,170
Average length	474.78	101.20	188.84	764.82
Average length with unique words	**38.98**	68.15	90.98	**198.11**

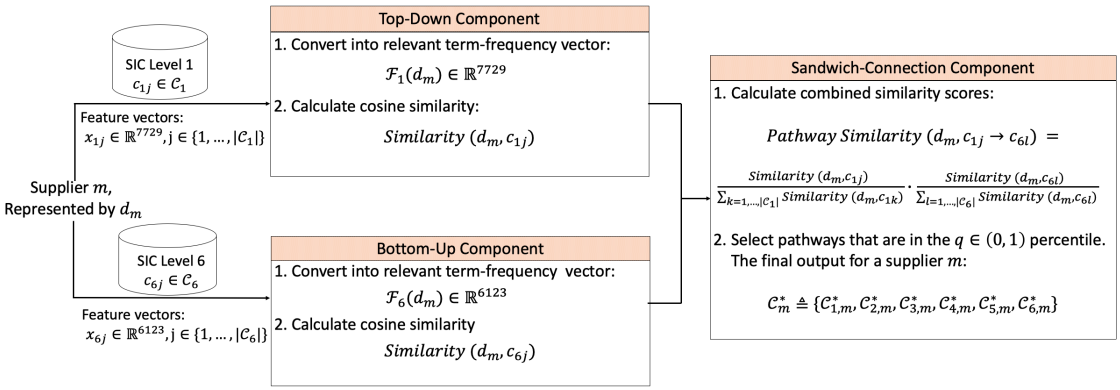
4.5 The Three-Component Classification Model

We now introduce a three-component hierarchical classification model, aiming to classify testing data (e.g., a suppliers' document d_m) into a *deep* and *broad* hierarchical taxonomy (e.g., an expanded SIC taxonomy). The model has three key components – the top-down classifier, the bottom-up classifier, and the sandwich-connection component, which are elaborated in Chapter 3.5.1. We discuss the challenges of gauging its accuracy and how we overcome them in Chapter 3.5.2. We report our model's accuracy relative to benchmark models for hierarchical classification in Chapter 3.5.3.

4.5.1 Methodology

The three-component classification model aims to predict a supplier m to its appropriate categories $c_{ij} \in \mathcal{C}_i$, for all levels $i \in \{1, \ldots, 6\}$. The model's three components are depicted in Figure 4.4. A supplier m, represented by its testing document d_m, is fed into both the top-down and bottom-up components of the model. In what follows, we provide a detailed description of each of these three components and their roles in the prediction. Also, we provide a detailed example for each component in Appendix 4.9.1.

Figure 4.4: The Classification Process in the Three-Component Model.



Top-Down Component.

The top-down component constructs a flat classifier that predicts a supplier m's testing document d_m to the appropriate level-1 categories $c_{1j} \in \mathcal{C}_1$. Although we illustrate using d_m, we allow for flexibility in the testing documents. That is, any combination of d_m^{gen}, d_m^{spe}, or d_m^{PO} can replace d_m. We will discuss the model performance under different data sources in Section 4.5.3.

The testing document d_m is first converted into the relevant term-frequency vector of the 7,729 informative features, represented by the function $\mathcal{F}_1(d_m) \in \mathbb{R}^{7729}$. Recall that each of the 21 level-1 categories c_{1j} has a corresponding *feature vector* $x_{1j} \in \mathbb{R}^{7729}$. We compare the similarity between a testing supplier's $\mathcal{F}_1(d_m)$ with each of the level-1 category feature vectors x_{1j}, for all $j \in \{1, ..., |\mathcal{C}_1|\}$ to determine the categories c_{1j} that the testing document d_m most closely belongs to. To reduce the bias against the size of the documents when creating term-frequency vectors and to handle sparse vectors (Maas et al. 2011), we employ cosine similarity:

$$\text{Similarity}(d_m, c_{1j}) = \frac{\mathcal{F}_1(d_m) \cdot x_{1j}}{||\mathcal{F}_1(d_m)|| \cdot ||x_{1j}||}. \tag{4.5.1}$$

Bottom-Up Component.

For the same testing supplier m, its testing documents d_m (or any combinations of d_m^{gen}, d_m^{spe}, or d_m^{PO}) is converted into a term-frequency vector in the dimension of 6,123 relevant informative features, so that $\mathcal{F}_6(d_m) \in \mathbb{R}^{6123}$. We compare its similarity with all feature vectors $x_{6j} \in \mathbb{R}^{6123}$ corresponding to each level-6 categories $c_{6j} \in \mathcal{C}_6$. Again, we employ the cosine similarity:

$$\text{Similarity}(d_m, c_{6j}) = \frac{\mathcal{F}_6(d_m) \cdot x_{6j}}{||\mathcal{F}_6(d_m)|| \cdot ||x_{6j}||}. \tag{4.5.2}$$

Sandwich-Connection Component.

After identifying the set of level-1 and level-6 categories that a supplier most likely belongs to, we examine the likely SIC pathways by "connecting the dots." This process takes advantage of the parent-child relationships. Among all 15,574 pathways $c_{1j} \to c_{6\ell}$, we can isolate the most likely pathways by examining the *product* of the normalized similarity scores:

$$\begin{aligned}&\text{Pathway Similarity}(d_m, c_{1j} \to c_{6\ell}) \\ &\quad\triangleq \frac{\text{Similarity}(d_m, c_{1j})}{\sum_{k=1,\ldots,|\mathcal{C}_1|} \text{Similarity}(d_m, c_{1k})} \\ &\qquad\times \frac{\text{Similarity}(d_m, c_{6\ell})}{\sum_{\ell=1,\ldots,|\mathcal{C}_6|} \text{Similarity}(d_m, c_{6\ell})}.\end{aligned}$$

The calculation of pathway similarity yielded similarity scores across all 15,574 pathways. We introduce a parameter $q \in (0, 1)$, which represents a percentile threshold within these similarity scores. Setting q to 0.99, for example, means that we retain only the top 1% of pathways based on their similarity scores. The F1 scores of the categorization model typically exhibit a unimodal distribution to q, peaking at an optimum $\hat{q}$. This is intuitive: selecting a lower q value causes the model to keep more pathways per supplier, boosting recall but reducing precision, whereas a higher q value keeps fewer pathways, enhancing precision at the expense of recall. Therefore, an intermediate q value that strikes a balance between precision and recall achieves the most accurate categorization model. We will test different $q \in (0, 1)$ to pinpoint the optimal $\hat{q}$ that yields the highest F1 scores in §4.5.3.

The sandwich-connection component combines the insights from the top-down and bottom-up components. For example, suppose that a supplier related to "boiler" is identified at level 6. If this supplier has a high level-1 similarity with "manufacturer" but a low similarity with "wholesale," then all pathways originating from the "wholesale" category would be pruned. Ultimately, our three-component classifica-

tion model predicts a supplier m into a set of pathways through level-1 to level-6 categories. In doing so, it automatically identifies all categories at the intermediate level ($i = 2, 3, 4, 5$). Thus, for a supplier m, it produces hierarchical classification:

$$\mathcal{C}_m^* \triangleq \{\mathcal{C}_{1,m}^*,\ \mathcal{C}_{2,m}^*,\ \mathcal{C}_{3,m}^*,\ \mathcal{C}_{4,m}^*,\ \mathcal{C}_{5,m}^*,\ \mathcal{C}_{6,m}^*\}.$$

where $\mathcal{C}_{i,m}^*$ denotes the set of categories $c_{ij} \in \mathcal{C}_i$ that the supplier m most likely belongs to at level i.

Observe that this classification output can be represented as a vector $\mathcal{C}_{i,m}^* \in [0,1]^{|\mathcal{C}_i|}$, where the j^{th} element is 1 if $c_{ij} \in \mathcal{C}_{i,m}^*$ and 0 if $c_{ij} \notin \mathcal{C}_{i,m}^*$. The overall classification $\mathcal{C}_m^*$ across all six levels can be represented by a binary vector $[0,1]^{|\mathcal{C}_1|+\cdots+|\mathcal{C}_6|}$ (in dimension 16,759). Thus, supplier m's classification output $\mathcal{C}_m^*$ can be considered as its "DNA" representation.

4.5.2 Overcoming the Challenges to Accuracy Evaluation

Gathering Partial True Labels for a Subset of Suppliers.

To evaluate the accuracy of our three-component classification model, we should evaluate its predictions about supplier m, $\mathcal{C}_m^*$, against its true labels $\mathcal{T}_m$,

$$\mathcal{T}_m = \{\mathcal{T}_{1,m},\ \mathcal{T}_{2,m},\ \mathcal{T}_{3,m},\ \mathcal{T}_{4,m},\ \mathcal{T}_{5,m},\ \mathcal{T}_{6,m}\},$$

where each $\mathcal{T}_{1,m}$ corresponds to the set of true labels in each level i.

As mentioned in §4.3.3, one key challenge is the absence of these actual label sets $\mathcal{T}$ for each supplier m. Furthermore, accurately identifying these sets for all 2,170 suppliers out of 15,574 possible true pathways is practically infeasible.

Therefore, we developed an approach that combines crowdsourcing with expert validation to gather true labels. The comprehensive details of this process for collecting true labels are provided in Appendix 4.9.2. Below, we provide an overview

of our method.

The procedure for collecting true labels includes five steps. The first two steps are designed to select a representative sample of suppliers, denoted by the subset $\mathcal{M}$, and to find the best estimate true pathways for this subset down to levels 4 or 5. This task was considered manageable because of the relatively smaller number of pathways at levels 4 and 5, which total $|\mathcal{C}_4 \cup \mathcal{C}_5| = 727$, as compared to $|\mathcal{C}_6| = 15,574$ at level 6. For the construction of subset $\mathcal{M}$, we employed the Pareto principle, selecting 68 suppliers responsible for 80% of the economic volume along with 210 others chosen at random from the remaining pool. This process yielded a subset size of $|\mathcal{M}| = 278$. So we estimate the true hierarchical categories for 278 suppliers down to levels 4/5:

$$\mathcal{T}_m^* = \{\mathcal{T}_{1,m}^*, \mathcal{T}_{2,m}^*, \mathcal{T}_{3,m}^*, \mathcal{T}_{4,m}^*, \mathcal{T}_{5,m}^*\}, \quad m \in \mathcal{M}.$$

The next three steps aim to identify the correct category at level 6. In our sample set of 278 suppliers, 107 (out of 727) unique categories were identified at level 4/5, which in turn encompassed 3,258 level 6 categories. Due to the substantial time and cost involved in obtaining true labels for a vast number of potential level 6 categories, we further narrowed down our focus to a smaller set of suppliers $\mathcal{M}' \subset \mathcal{M}$. To do so, in Step 3, we counted the number of suppliers in each true level 4/5 category, and selected the top four level 4/5 categories (see Table 4.12 of Appendix 4.9.2). These top four categories include 105 unique suppliers from Cranswick, thus setting the size of $\mathcal{M}'$ as 105 (i.e., $|\mathcal{M}'| = 105$). For $\mathcal{M}'$, we have repeated the procedure of crowdsourcing and expert validations to further identify the best estimate of the true labels at level 6, so that we have

$$\mathcal{T}_m^* = \{\mathcal{T}_{1,m}^*, \mathcal{T}_{2,m}^*, \mathcal{T}_{3,m}^*, \mathcal{T}_{4,m}^*, \mathcal{T}_{5,m}^*, \mathcal{T}_{6,m}^*\}, \quad m \in \mathcal{M}'.$$

These best estimates of the true labels will be used to gauge the accuracy of the three-component classification model in §4.5.3.

Accuracy Metric.

For a supplier $m \in \mathcal{M}$, the three-component model predicts the set of categories $\mathcal{C}^*_{i,m}$ at level i. We want to measure the accuracy of this prediction with respect to the best estimate of the true labels, $\mathcal{T}^*_{i,m}$. Observe that each set may contain multiple elements, i.e., $|\mathcal{C}^*_{i,m}| \geq 1$ and $|\mathcal{T}^*_{i,m}| \geq 1$, leading to multiple predicted and true pathways.

Let $TP_{i,m}$ (True Positive) denote the number of categories that are correctly predicted (i.e., $c_{ij} \in C^*_{i,m}$ when $c_{ij} \in \mathcal{T}^*_{i,m}$), and $FN_{i,m}$ (False Negative) denote the number of categories that should have been predicted but were not (i.e., $c_{ij} \notin C^*_{i,m}$ when $c_{ij} \in \mathcal{T}^*_{i,m}$). Formally, we have

$$TP_{i,m} \equiv |\mathcal{C}^*_{i,m} \cap \mathcal{T}^*_{i,m}|, \quad FN_{i,m} \equiv |\overline{\mathcal{C}^*_{i,m}} \cap \mathcal{T}^*_{i,m}|.$$

A higher number of predicted categories (i.e., larger $|C^*_{i,m}|$) can increase the likelihood of correctly identifying one or more true labels, but also raise the possibility of wrong predictions. To assess the accuracy of the classification model, we utilize the precision and recall metrics:

$$\text{Precision}_{i,m} \equiv \frac{TP_{i,m}}{|\mathcal{C}^*_{i,m}|}, \quad \text{Recall}_{i,m} \equiv \frac{TP_{i,m}}{TP_{i,m} + FN_{i,m}}.$$

The former metric measures the proportion of accurately predicted labels to the total number of categories predicted; while the latter metric measures the proportion of correctly predicted labels to all the categories that should have been predicted. Perfect precision indicates that every predicted category is correct, but it does not guarantee that all correct categories have been predicted. Perfect recall, on the

other hand, ensures that all correct labels are predicted, but it does not specify the number of predicted categories.

To balance the precision and recall metrics, we employ the widely-used F1 score (Holden and Freitas 2006, Costa et al. 2007). At hierarchical level i, the F1 score for an individual supplier m and for the average across all M suppliers are respectively,

$$F1_{i,m} \equiv \frac{2 \cdot \text{Precision}_{i,m} \cdot \text{Recall}_{i,m}}{\text{Precision}_{i,m} + \text{Recall}_{i,m}},\ F1_i = \frac{1}{M} \sum_{m=1}^{M} F1_{i,m}.$$

4.5.3 Accuracy of the Three-Component Classification Model

We run the classification of the entire 2,170 suppliers. Using the true label estimate gathered, we are able to estimate the classification accuracy down to levels 4/5 for the suppliers $m \in \mathcal{M}$, and down to level 6 for suppliers $m \in \mathcal{M}'$. These accuracy results are deemed representative of the accuracy of the classification of entire $M = 2,170$ suppliers.

We first demonstrate the three-component model's superior accuracy over benchmark models and how its structure and data flexibility impact the classification accuracy. We then examine how our each of our model components reflects the procurement expert's know-how to improve accuracy.

Classification Accuracy Evaluation.

First, we tested with a range of q values to determine the optimal $\hat{q}$, which strikes a balance between precision and recall, thereby optimizing the accuracy of the categorization model. We found the optimal value of q, denoted as $\hat{q}$, to be 0.99.

Then, under the optimal value $\hat{q}$, we compare the performance of our three-component model versus two benchmark hierarchical classification models. The first benchmark is the traditional top-down model that predicts level-by-level along the SIC pathway (Dumais and Chen 2000, Cesa-Bianchi et al. 2006, Esuli et al. 2008,

Cerri et al. 2014), and the second benchmark is the traditional bottom-up model that predicts level 6 and employs heuristic pruning (Ceci and Malerba 2007). See Appendix 4.9.4 for a detailed explanation of benchmark models.

Table 4.5: Average F1 Score in Three-Component Model and Benchmark Models. Here, F1 at levels 1-5 are aggregated with 278 suppliers in the set $\mathcal{M}$, and level 6 is aggregated with 105 suppliers in the set $\mathcal{M}'$. In all models, $q = 0.99$.

	Three-Component Model		Benchmark Models			
			Top-down		Bottom-up	
	d_m	flex	d_m	d_m^{gen}	d_m	$d_m^{spe} \cup d_m^{PO}$
	(1)	(2)	(3)	(4)	(5)	(6)
Level 1	0.858	**0.935**	0.802	**0.849**	0.642	**0.663**
Level 2	0.569	**0.675**	0.561	**0.579**	0.403	**0.421**
Level 3	0.449	**0.530**	**0.438**	0.387	0.321	**0.342**
Level 4	0.382	**0.457**	**0.356**	0.294	0.283	**0.306**
Level 5	0.377	**0.449**	**0.342**	0.277	0.281	**0.305**
Level 6	0.367	**0.424**	**0.204**	0.158	0.230	**0.301**

Table 4.5 presents the $F1_i$ scores for $i \in \{1, 2, 3, 4, 5, 6\}$ of the three-component model and two benchmark models for different combinations of supplier data d_m (d_m^{gen}, d_m^{spe}, d_m^{PO}). For the top-down benchmark model, the prediction accuracy drops significantly with deeper levels due to error propagation. That is, if the model's prediction is incorrect at level 1, the prediction will also be incorrect at level 2, and so on. The benchmark bottom-up performs better than the top-down model at deeper levels (e.g., levels 5-6), but is worse at predicting the high levels because it lacks the contextual information.

Our three-component model utilizes the advantages of two benchmark models. Comparing the columns with the same input suppliers' information d_m (i.e., columns 1, 3, 5), we observe that the three-component model outperforms benchmark models at every hierarchical level $i \in \{1, 2, 3, 4, 5, 6\}$. This suggests that our three-component model structure contributes to the improvement in classification accuracy.

Furthermore, we analyzed various data combinations (all combinations of d_m^{gen}, d_m^{spe}, d_m^{PO}) for each model. We found that the top-down model generally performs better when employing d_m^{gen} (i.e., column 4) instead of all combined texts d_m (i.e., column 3); while the bottom-up model tends to perform better when employing $d_m^{spe} \cup d_m^{PO}$ (i.e., column 6) instead of all combined text data (i.e., column 5). This indicates that the additional text contained in d_m results in noise and hinders the classification. For the three-component model, optimal accuracy was achieved by flexibly applying d^{gen} to its top-down component and $d_m^{spe} \cup d_m^{PO}$ to its bottom-up component always provided the highest accuracy (i.e., column 2). This suggests that when additional data flexibility is allowed in the three-component model, the classification accuracy is significantly superior to those of the benchmarks. In particular, its level-6 accuracy is comparable to the benchmarks' level-2 or level-3 accuracy levels.

In sum, the three-component model improves prediction accuracy by (1) taking advantage of the prediction capabilities of top-down and bottom-up components, and (2) providing the flexibility to incorporate different sources of suppliers' information (i.e., supplier contextual description, product descriptions, and supplier actual procurement) to mitigate the effects of noise in the input data.

Note that the accuracy evaluation in Table 4.5, each supplier is treated with equal weight, irrespective of the monetary value of transactions per supplier. This choice is aimed at maintaining consistency and simplicity in the evaluation metrics. We acknowledge that transactions of differing financial magnitudes from individual suppliers might warrant distinct consideration. Future studies could explore weighting schemes that account for these factors to refine the analysis.

Three-Component Model Structure and Procurement Expert's Know-How.

In the three-component model, recall that the top-down component constructs a flat classifier to predict level-1 categories $c_{1j} \in \mathcal{C}$. It is designed to process general

descriptions and reflects the procurement expert's contextual knowledge. In contrast, the bottom-up component is designed to focus on specific word matching to predict level-6 categories $c_{6j} \in \mathcal{C}$, which reflects the procurement expert's familiarity with specific product terms and word associations. The sandwich-connection component actively utilizes the parent-child relationships in the hierarchical taxonomy reflecting the procurement expert's logic when applying classification. We next investigate how this structure reflects the procurement expert's know-how to improve classification accuracy against the benchmark methods.

Table 4.6: Comparative F1 Scores at Level 1 for 278 Suppliers Using Different Classification Models and Different Input Texts. Note that F1 at level 1 is aggregated with 278 suppliers. In all models, $q = 0.99$.

	General description	Specific descriptions			All texts
F1 at level 1	d_m^{gen} (1)	d_m^{spe} (2)	d_m^{PO} (3)	$d_m^{spe} \cup d_m^{PO}$ (4)	d_m (5)
Benchmark Top-down model	**0.849**	0.746	0.610	0.742	0.802
Benchmark Bottom-up model	0.563	0.638	0.619	**0.663**	0.642
Three-Component model (same data source for both entries)	**0.898**	0.849	0.723	0.848	0.858
Three-Component model (best data source for each entry)	**0.935**				

Table 4.6 shows the F1 scores at level 1. As previously discussed, the benchmark top-down model performs better than the bottom-up model in predicting level-1 categories. We observe that the three-component model brings visible improvement in level-1 accuracy compared to the benchmark bottom-up model. This demonstrates the importance of incorporating a top-down component in predicting higher-level categories. Specifically, by doing so, the classification accuracy improved from the bottom-up model's accuracy of 0.663 to 0.848 when classifying with specific product descriptions ($d_m^{spe} \cup d_m^{PO}$). Moreover, by being able to "connect the dots," we improve from the top-down model's accuracy of 0.849 to 0.898 when classifying with general product descriptions (d_m^{gen}). Finally, by incorporating data flexibility to reduce noise, the three-component model improved the accuracy score to 0.935.

Table 4.7: Comparative F1 Scores at Level 6 for 105 Suppliers Using Different Classification Models and Different Input Texts. Note that F1 at level 6 is aggregated with 105 suppliers. In all models, $q = 0.99$.

	General description	Specific descriptions			All texts
F1 at level 6	d_m^{gen} (1)	d_m^{spe} (2)	d_m^{PO} (3)	$d_m^{spe} \cup d_m^{PO}$ (4)	d_m (5)
Benchmark Top-down model	0.158	0.190	0.138	0.189	**0.204**
Benchmark Bottom-up model	0.242	0.246	0.274	**0.301**	0.230
Three-Component model (same data source for both entries)	0.349	**0.368**	0.315	0.359	0.367
Three-Component model (best data source for each entry)	**0.424**				

Table 4.7 presents the F1 scores for level 6, showing that the benchmark bottom-up model exceeds the performance of the benchmark top-down model at this level, across various input texts from Columns (1) to (5). Also, our three-component model significantly enhances the accuracy at level 6, demonstrating the effectiveness of including a bottom-up component for predicting more specific categories. Specifically, the best performance of the benchmark top-down model is achieved with the use of comprehensive texts in Column (5) (i.e., d_m), where the three-component model improves classification accuracy from 0.204 to 0.367 with the same text. The best performance of the benchmark bottom-up model is achieved with the specific product descriptions in Column (4) (i.e., $d_m^{spe} \cup d_m^{PO}$), where the three-component model improves classification accuracy from 0.301 to 0.359. Finally, by incorporating data flexibility, the three-component model improved the level-6 accuracy score to 0.424.

4.6 Decision Support Tools

The methodology thus far has organized vast text data and purchase order data from 2,170 suppliers into categories $c_{ij} \in \mathcal{C}_i$ of a hierarchical taxonomy that is both deep (6 levels) and broad (15,574 leaf nodes). The classification helped make sense of the vast amount of purchase order records in the form of unstructured text data.

In this section, we present the complementary decision support tool that can help its users in converting the classification results into opportunities for savings.

The aim of conducting a spend analysis is to identify opportunities for savings and recommend target suppliers to initiate the RFQ process to negotiate lower costs. Implementing the RFQ ranges from holding private negotiations with the suppliers to designing and holding public auctions. The outcome of a successful RFQ usually involves switching suppliers and managing new relationships which entails significant commitment of the manufacturer's internal resources (and may sometimes require re-structuring parts of its procurement processes). Thus, an RFQ recommendation must present convincing evidence of the potential cost savings. We next describe how our decision support tool offers insights into the supplier and product categories the buyers should target to seek price/volume discounts (§4.6.1), and offers tools for easy cross-comparisons of many suppliers to understand the nuances in the procurement practice (§4.6.2).

4.6.1 Identify Leverage Categories

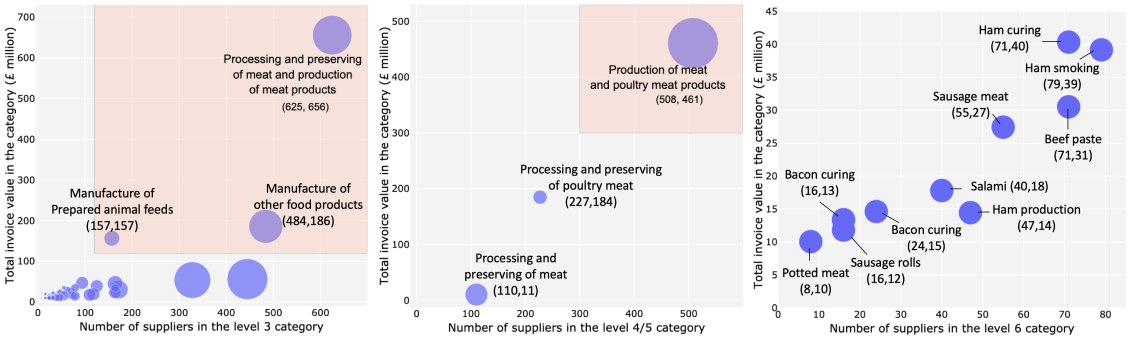
Recall that the three-component classification model produces for each supplier m, a prediction $\mathcal{C}_m^*$ or its classification "DNA." Utilizing this information along with the invoice value, our decision support tool helps to identify the *leverage* supplier categories. They are the class of suppliers that pose low supply risk to the manufacturer but whose products have a high impact on the manufacturer's costs. These suppliers are where the cost-savings or RFQ opportunities generally arise from because the buyers hold the dominant position in the buyer-supplier relationship.[3]

It does so by performing a widely applied method for strategic sourcing called

[3]In contrast, if a supplier provides a unique product or service that not many others can provide (e.g., Foxconn) that is critical to a buyer (e.g., Apple), the buyer is exposed to supply risk that can significantly impact its profits. The buyer must manage such "high risk / high profit impact" suppliers delicately, and they should not be a target for issuing RFQs.

Kraljic analysis (Kraljic 1983, Webb 2017). It segments suppliers into four types based on their contribution to the buyer's risk and profit. The number of suppliers in the category is a good proxy for supply risk (the higher the number of similar suppliers, the lower the risk); and the total spend on a product category is a good proxy for the impact on a buyer's profit (the higher the total invoice value, the higher the impact on profit). The output of our classification model allows for easy identification of the suppliers in these leverage categories $c_{ij} \in \mathcal{C}_i$ at any hierarchy level $i \in \{1, 2, 3, 4, 5, 6\}$.

Figure 4.5: Illustration of Kraljic Analysis to Identify the Leverage Categories at Level 3 (left), Level 4/5 (mid), and Level 6 (right). Note that since a supplier can be categorized into multiple nodes in a hierarchy, the sum of the suppliers can exceed 2,170. $q = 0.99$.



The hierarchical classification of suppliers facilitates seamless navigation through various supplier and product categories, helping in the identification of key leverage points. Figure 4.5 shows the categories in hierarchy level 3 (left panel), level 4/5 (middle panel), and level 6 (right panel) for Cranswick plc's supplier/product categories. Presenting each level is very useful for strategic and organisational reasons. For example, in large-scale procurement settings, there are dedicated department or teams for managing different supplier relationships. A department that handles animal feed suppliers would differ from one managing meat product suppliers, and within a department, different teams could be dedicated to poultry and sausage meat. This level of detail, as illustrated in Figure 4.5, supports targeted and ef-

fective management within the organization. The x-axis denotes the number of suppliers per category, the y-axis shows the total invoice values for all suppliers within a category, and the bubble size indicates the product diversity within the category or the count of sub-categories.

The left panel of Figure 4.5 illustrates the categories in level 3. We can identify three leverage categories positioned on the top-right: "Manufacture of other food products", "Processing and preserving of meat and production of meat products" and "Manufacture of prepared animal feeds". Zooming into the category "Processing and preserving of meat and production of meat products," we see its level-5 sub-categories, as illustrated in the middle panel of Figure 4.5. We observe that it consists of three detailed supplier categories: "Production of meat and poultry meat products", "Processing and preserving of poultry meat" and "Processing and preserving of meat." The right panel further narrows down to reveal product categories within "Production of meat and poultry meat products," showing that they can be distinguished by specific product types, such as "ham curing" and "sausage meat". Note that level-6 represents the most detailed classification with no sub-categories, all bubbles appear uniform in size.

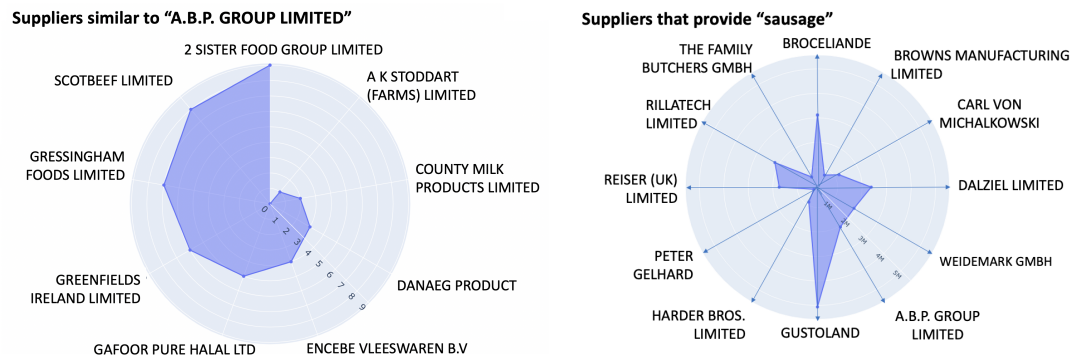
4.6.2 Comparison of Suppliers

For the identified leverage categories, the decision support tool enables cross-comparisons of their suppliers and product categories to help in developing a fine-grained picture of the supplier-product relationships. One can compute the similarity between any two suppliers m_1 and m_2 by comparing their categorizations (or DNA's) $\mathcal{C}^*_{m_1}$ and $\mathcal{C}^*_{m_2}$, via cosine similarity or weighted difference of their DNA vectors.

In the left panel of Figure 4.5, there were 625 suppliers categorized under level-3 as "Processing and preserving of meat and production of meat products." Within this category, we can list those who compete with "A.B.P. GROUP LIMITED" based

on their similarities. This is illustrated in the left panel of Figure 4.6, which identifies "2 SISTER FOOD GROUP LIMITED" emerges as the most closely matched competitor. Furthermore, our decision support tool is capable of listing suppliers not just by their past supply records but also by potential supply capabilities. The right panel of Figure 4.6 demonstrates this by ranking suppliers that are capable of providing "sausage meat," arranged according to their total invoice values. This figure offers valuable insights that are not easily obtained through manual methods.

Figure 4.6: Similar Suppliers with a Focal Supplier (left) and Similar Suppliers with a Focal Product (right).



In the right panel of Figure 4.5, there were 71 suppliers within the product category "Ham curing." We found that "Supplier 383" has a significantly higher amount of invoice value than all of its competitors combined. Moreover, "Supplier 1390" and "Supplier 383" provide a similar range of products but Supplier 1390 supplies less than a tenth of the invoice value as Supplier 383 and also does not supply other products such as "sausage meat" or "bacon curing." Increasing the purchase variety and volume from Supplier 1390 would create competitive pressures that can help lower purchase costs for the buyer in this product category. Such information can be utilized to inform the development of RFQ recommendations and realize its savings potential.

4.7 Advantage of Automated Spend Analysis: A Simulation Study

The automation discussed thus far enhances classification by providing an initial, high-fidelity categorization. It is designed to complement, rather than replace, manual input. There are three key reasons why automation is beneficial: it expands scope, improves accuracy, and boosts adaptability by enabling more frequent analysis. First, automation enables a manufacturer to analyze all its suppliers rather than just a subset, thereby expanding the *scope* of spend analysis. Second, automation improves the *accuracy* of classifications. Instead of creating classifications from scratch, the manufacturer can focus its efforts on reviewing and correcting any errors in the initial automated classification. Third, automation significantly increases the speed of spend analysis. With automated processes, what once took months can now be accomplished in a matter of days. This efficiency allows firms to conduct spend analysis more regularly, enabling them to respond more swiftly to market conditions. The improved *adaptability* that comes with more frequent analysis helps firms stay competitive and make timely adjustments to their strategies based on the latest data.

In this section, we estimate the savings achieved through the automation of a companys spend analysis. To do this, we conducted a simulation study, introducing a model of Cranswick plc's supply chain and calibrating it with the provided data. We differentiate between the savings generated by automation and those typically realized through manual analysis alone.

4.7.1 Simulation Experiment Design

Model of Cranwick's Supply Chain.

Cranswick interacts with M suppliers to order N different level-6 product categories. Specifically, Cranswick plc's supply chain includes $M = 2,171$ suppliers and $N = 3,258$ product categories, as determined by examining that $|\mathcal{C}^*_{6,m}| = 3,258$. The distribution of which suppliers provide which product categories and the corresponding economic value of these supplies can be represented by two random M-by-N matrices, S and P.

The matrix S is a binary matrix that represents the supplier-product relationship of Cranswick plc. If supplier m can supply product category n, then its element $s_{mn} = 1$; otherwise the element $s_{mn} = 0$. Examining the classification $\mathcal{C}^*_m$, for all m, we find that each supplier on average supplied 5.03 level-6 product categories, with a minimum of 1 and maximum of 13. Thus, to construct this supplier matrix S, for each supplier m, we assign a random product scope (i.e., the number of products from a supplier) characterized by a binomial distribution with parameters $n = 13$ and $p = 5.03/13$. We then randomly select the corresponding number of product(s) from the set of N products.

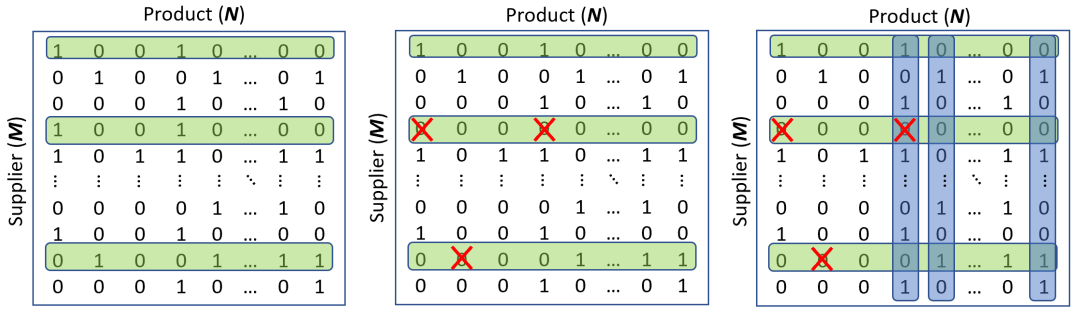
The matrix P represents the purchase order p_{mn} that a supplier m charge when supplying a certain quantity of product category n. Recall from Table 4.1, we had examined 556,866 purchase orders over a two year period totaling £1.57bn. These purchase orders were from 2,171 unique suppliers, and thus each supplier had an average an invoice value of £723,776 over two year period, which translates into an average annual invoice of £361,888 per supplier. Given that suppliers supplied on average 5.03 level-6 product categories, for a single product category the average invoice value per product category p_{mn} is £72,377. Thus, for each product-level category, we model the invoice value as an exponential distribution, with CDF,

$F(p) = 1 - e^{-p/72377}$.

Model of Spend Analysis: Scope, Accuracy, and Adaptability.

Next, we explain how we assess the *scope*, categorization *accuracy*, and *adaptability* of spend analysis for both manual and automated spend analysis. To illustrate this, we use Figure 4.7, which utilizes the supplier-product matrix S.

Figure 4.7: Model of Scope and Accuracy of Spend Analysis, as well as Kraljic Analysis on Supplier-Product Matrix S.



Scope. Manual spend analysis and automated spend analysis differ primarily in their supplier scope. In manual spend analysis, procurement experts typically focus on a small subset of suppliers that represent large proportion of the total spend. For example, in the case of Cranswick plc, 68 out of 2,171 suppliers are responsible for 80% of the total invoice value. For our normalization, we assume that automated spend analysis can examine all of suppliers corresponding to 100% of the total spend, while manual spend analysis covers only a subset of suppliers that represent 70-90% of the total spend. For illustration, the left panel of Figure 4.7 shows the scope of analysis. The highlighted rows represent the scope for manual analysis, whereas all rows represent the scope for automated spend analysis.

Accuracy. Within the given scope of analysis, manual and automated spend analysis differ in the accuracy of suppliers' classification. In automation-aided spend analysis, procurement experts are provided with an initial categorization of suppliers, which they manually check and amend as necessary. In contrast, manual

analysis requires the firm to conduct the initial categorization from scratch, which takes more time and effort and has a greater risk of misclassification. Thus, with automation, classification tends to be more accurate than when it is performed via manual methods alone. In our simulation, to represent the level of accuracy, we will randomly misclassify the suppliers by altering the 1's and 0's as illustrated in the mid panel of Figure 4.7. The elements s_{mn} marked with "×" correspond to misclassified supplier-product links. Manual spend analysis results in an accuracy between 70%-90%, meaning 10%-30% of columns will be marked with ×" in the highlighted rows. In contrast, for automation-aided spend analysis, we normalize the accuracy to be 100% (i.e., none of the columns will have "×").

Adaptability. One of the key benefits of automating spend analysis is the ability to perform it quickly and cost-effectively. This allows for frequent analysis, enabling better adaptability to changing market conditions, including fluctuations in suppliers, products, volumes, and prices. To model this effect, we consider the years 2019 and 2020. We assume that manual spend analysis can only be conducted once during this period, while automated spend analysis can be performed once in 2019 and once in 2020.

Model of Saving Realization.

Once the suppliers are categorized, the procurement team must (i) identify savings opportunities by pinpointing leverage suppliers and (ii) conduct a request-for-quote (RFQ) process. Next, we describe how these steps are modeled and calibrated in our simulation.

Identification of Saving Opportunities. Recall that to identify savings opportunities, a Kraljic analysis is conducted to identify product categories with high spend volume and a large number of suppliers. In our simulation model, these product categories can be identified by ordering the values of $\sum_m p_{mn}$ and $\sum_m s_{mn}$ respectively. These categories are illustrated by the highlighted columns in the

supplier-product matrix S in the right panel of Figure 4.7.

Implementing RFQ and the Estimated Cost Savings. Once the target level-6 product categories are identified, Cranswick plc will initiate an RFQ for all the suppliers in the product categories. The RFQ can involve either a private negotiation with an individual supplier or a public auction. Initiating RFQs can be a costly process, so they would typically do so only if they expect a savings of between 5-10% of the current invoice values.[4] Thus, for each supplier identified in the leverage categories, we apply a random discount δ, which is modeled as the uniform distribution between 5% and 10%, to set the new invoice value $p_{mn} = \delta \cdot p_{mn}$.

Overall Savings from Manual Spend Analysis. Implementing RFQs on all selected leverage product categories recommended by manual spend analysis typically translates into 2-3% in overall procurement cost savings in a successful industry practice. To reflect this percentage, we calibrate the Kraljic analysis to identify the product categories that are in the top 33% for economic volume *and* the top 33% for the number of suppliers. In examining the effect of adaptability in the second period, we produce another randomized invoice matrix P to reflect the changes in the market. However, to reflect the lasting benefit of spend analysis from the first period, we will employ lower average invoice prices (e.g., 2-3% lower than č72, 377 of the first stage).

4.7.2 Simulation Results

In what follows, we will observe that automation of spend analysis creates *additional* savings over the current manual spend analysis by enabling the analysis of a greater scope of suppliers with increased categorization accuracy. Moreover, by enabling spend analysis to be performed more frequently, automation allows these benefits to compound over time.

[4]The 5-10% is the range of realized values that are validated by our industry partners based on their experiences in the industry.

In each simulation iteration, the model randomly selects the matrices S and P according to the calibrated distributions and performs nine manual spend analyses. These analyses combine different scopes of supplier coverage (70%, 80%, and 90% of invoice value) with varying classification accuracies (70%, 80%, and 90%). Additionally, automated spend analysis is conducted with both scope and accuracy normalized to 100%. For each combination of these parameters, we conducted 1,000 iterations.

Table 4.8: Comparative Results of Manual vs. Automated Spend Analysis Across Varying Scopes and Accuracies (1-Year)

	Manual Analysis (1-year)			Automated Analysis (1-year)	
Cost (£million)	Manual Scope	Manual Accuracy	Manual Saving (%)	Auto Saving (%) (100% Scope & 100% Accuracy)	%-Δ(Auto-Manual) (£million)
785	70%	70%	2.25%	3.43%	+1.18% (+9.26)
796	70%	80%	2.42%	3.37%	+1.28% (+10.20)
786	70%	90%	2.50%	3.38%	+0.88% (+6.92)
780	80%	70%	2.33%	3.35%	+1.02% (+7.88)
783	80%	80%	2.54%	3.37%	+0.83% (+6.50)
792	80%	90%	2.76%	3.33%	+0.57% (+4.51)
783	90%	70%	2.44%	3.27%	+0.83% (+6.50)
795	90%	80%	2.76%	3.38%	+0.62% (+4.93)
795	90%	90%	3.00%	3.43%	+0.43% (+3.42)

Table 4.9: Comparative Results of Manual vs. Automated Spend Analysis Across Varying Scopes and Accuracies (2-Year)

	Manual Analysis (2-year)			Automated Analysis (2-year)	
Cost (£million)	Manual Scope	Manual Accuracy	Manual Saving (%)	Auto Saving (%) (100% Scope & 100% Accuracy)	%-Δ(Auto-Manual) (£million)
1579	70%	70%	2.25%	5.07%	+2.82% (+44.53)
1600	70%	80%	2.42%	5.00%	+2.58% (+41.28)
1577	70%	90%	2.50%	4.99%	+2.49% (+39.27)
1570	80%	70%	2.33%	4.99%	+2.66% (+41.76)
1565	80%	80%	2.54%	5.00%	+2.46% (+38.50)
1587	80%	90%	2.76%	4.93%	+2.17% (+34.43)
1578	90%	70%	2.44%	4.91%	+2.47% (+38.97)
1575	90%	80%	2.76%	5.03%	+2.27% (+35.75)
1583	90%	90%	3.00%	5.02%	+2.02% (+31.98)

Table 4.8 presents the results for the 1-year analysis, while Table 4.9 presents the results for the 2-year analysis. From both tables, we observe that the total invoice

cost from all suppliers across all product categories ranges between £780 million and £796 million annually (See Table 4.8), and over a two-year period, the total cost ranges from £1.565 billion to £1.600 billion (See Table 4.9). These figures align with Cranswick plc's annual procurement expenditure, as indicated in Table 4.1.

In Table 4.8, manual spend analysis achieves annual savings ranging from 2.25% to 3.00%, consistent with the RFQ implementation benchmarks observed in the industry. As expected, we observe that increased scope and accuracy in manual spend analysis contribute to higher savings. Examining the value of automated spend analysis, we find that it leads to overall annual savings of 3.27% to 3.43%. Compared to manual spend analysis, this results in an additional 0.43% to 1.28% savings (equivalent to £3.42 million to £10.20 million) in the first year.

In Table 4.9, the automated spend analysis generates savings between 4.91% and 5.07% over a two-year period, providing an additional 2.02% to 2.82% compared to manual spend analysis over the same duration. This corresponds to an *additional* £31.98 million to £44.53 million in savings over two years, or an extra £16 million to £22 million annually compared to manual spend analysis.

4.8 Discussion

To the best of our knowledge, our methodology is the first academic work to formalize the automation of spend analysis using NLP and machine learning. We highlight its potential contributions to a path towards the evolution of Industry 4.0 (Olsen and Tomlin 2020).

4.8.1 Impact on Procurement Practice

Our methodology has the potential to democratize access to spend analysis to many small and medium-sized enterprises (SMEs). The supply chain complexity of many SMEs can be comparable to that of large firms. However, due to their volume of

purchases being smaller, the estimated value of potential savings from conducting a spend analysis often does not justify the cost of hiring multiple procurement consultants over an uncertain prolonged duration. Automation of spend analysis removes these costs and makes accurate spend analysis accessible.

For example, our methodology has also been applied to conduct a spend analysis for a mid-cap company in the industrial sector. A private equity firm who had recently acquired it wanted to estimate the potential value improvement (e.g., by restructuring the supply chain) and provided us with the company's raw procurement data for 2020. The complexity of data was comparable to that from Cranswick plc, and consisted of 86,629 purchased orders, 25,025 products, and 1,829 suppliers. However, the total annual procurement spend was significantly lower at approximately £80 million. Our methodology identified four "leverage categories," whose combined value sums to roughly £22 million per year. Utilizing the decision tools, we were able to generate a list of supplier targets to recommend RFQs, which we estimated would translate into approximately 1.5-2.5% of the overall cost (approx. £2 million) in annual savings if implemented.

Since the manual spend analysis was previously not accessible for such firms, the value of the automated spend analysis would directly correspond to the savings that could be achieved. Thus, automated spend analysis could permeate throughout the industrial sectors. For example, a digital platform that provides automated spend analysis once buyers upload their transaction records can be offered, which would enable them to monitor their spending in real-time. Such developments would further accelerate the automation of spend analysis across different industries and transform the way in which we monitor how physical "things" are produced and distributed.

4.8.2 Generalizeability of Methodology

Our paper introduces a methodology that relaxes the reliance on extensive datasets commonly needed for machine learning algorithms, and instead train small but informative data efficiently. This approach enhances the flexibility of our methodology in manufacturer settings. For example, our methodology has been applied in a merger and acquisition setting where a German industrial manufacturer acquired a Swedish company. The German acquirer had a detailed internal supplier-classification database and taxonomy and wanted to classify the Swedish firm's extensive supplier list according to their existing system. The German company shared with us their own hierarchical taxonomy and their supplier database. Instead of using SIC taxonomy as the training data (as shown in "Utilizing 'Small' Data" in Figure 4.1), we utilized the German company's hierarchical database to train our three-component model. Our three-component model was then able to provide the classification of the Swedish firm's suppliers according to the German company's taxonomy, facilitating its timely integration.

Also, our methodology takes large sets of unstructured data and converts them into a structured format that can provide strategic insights. This feature allows for applicability across various industries beyond manufacturing, particularly where structured data is limited but abundant descriptive documentation is available. For example, in the financial or legal services sectors, there are extensive documentations of regulation, codes of practice, and compliance reports. Our three-component model may map these processes into a hierarchical structure. Once the hierarchical structure of the regulations, sections, clauses, and sub-clauses are understood, the three-component model could be trained on the extensive regulatory documentation. It could then classify new documents (e.g., live cases) into the relevant categories within the regulatory framework. After categorization, the model pinpoints which sub-clauses or sections are most frequently associated with live cases. Such classifica-

tion supports decision-making by highlighting areas that require attention, allowing for proactive measures to address compliance issues or service needs.

While we recognize the growing importance and capabilities of large language models (LLMs) in various applications, LLMs often struggle with categorization tasks requiring specific industry knowledge. LLMs must be supplemented with large, industry-specific datasets. Thus, an LLM is not a substitute, but a complement of our methodology that can help improve it further. For instance, rather than relying on static SIC documentation as we currently do, LLMs could be employed to incorporate the ability to *learn* industry-specific settings and update detailed databases of suppliers and product information. How to incorporate the capabilities of generative AI to complement our method may lead to fruitful directions for research and development.

4.8.3 Conclusion

This study has introduced a solution for automating the spend analysis process by leveraging large-scale industrial procurement data. Our findings illustrate that the developed hierarchical classification model, consisting of three components, successfully categorizes documents into any hierarchical taxonomy with high accuracy. The model outperforms current benchmarks, showing particular strength in handling taxonomies that are both deep and broad, as often required in real-world scenarios. Furthermore, the model's design to accommodate various types of textual data (e.g., general vs. specific) enhances its precision in classification. We have effectively demonstrated that spend analysis can be automated, incorporating the expertise of procurement professionals even in the absence of large datasets and precise supplier labels. This proof of concept paves the way for further research and development initiatives that could revolutionize the practice of spend analysis.

Due to the vast scale of the digital infrastructures in large manufacturing firms

and their links to people and processes, a drastic change is prohibitively costly and risky, and often infeasible within a reasonable time frame. Similar to re-purposing and re-connecting existing physical infrastructures of a large housing complex (e.g., pipes and wires), we presented a methodology that utilizes existing digital infrastructures by gathering them and generating insights. Such a solution represents the spirit of gradual process improvement (Fine and Porteus 1989) that is necessary to evolve towards Industry 4.0. While data analytics involving NLP and machine learning, and artificial intelligence more generally, is a burgeoning field, its reach has been comparatively limited in many of the industrial sectors, such as large-scale procurement. We believe that many other B2B operations management contexts are currently untapped by data analytics. For example, an important social agenda is for manufacturers to reduce their carbon emissions. While firms are making important strides in reducing their direct (Scope-1) and indirect (Scope-2) carbon emissions, significant challenges remain in addressing emissions in their supply chain (Scope-3), which represent the vast majority of emissions. Our research method enhances transparency in the supply chains of the firms, and could pave ways for them to accurately track and manage their Scope-3 carbon emissions. We hope that operations management scholars can help lead the effort to modernize the industrial process.

4.9 Appendix

4.9.1 Detailed Explanation for Three-Component Model

Recall that in Section 4.4, each level-1 category $c_{1j} \in \mathcal{C}_1$ ($j \in \{1, ..., |\mathcal{C}_1|\}$) was represented by a feature vector $x_{1j} \in \mathbb{R}^{7729}$, and each level-6 category $c_{6j} \in \mathcal{C}_6$ ($j \in \{1, ..., |\mathcal{C}_6|\}$) was represented by a feature vector $x_{6j} \in \mathbb{R}^{6123}$.

Top-Down Component

First, we convert the testing supplier m's text document d_m into the relevant term-frequency vector, denoted by $\mathcal{F}_1(d_m) \in \mathbb{R}^{7729}$. To do so, we extract the unigrams and bigrams from d_m, restrict the extracted terms to the 7,729 selective informative features, and count the frequency of the overlapped terms. For example, suppose d_m = *"We manufacture chicken wing chicken thigh."* We extract ten terms (five unigrams and five bigrams). Say, only five terms (e.g., "manufacture,""chicken,""wing,""thigh," and "manufacture chicken") are overlapped with level-1 features. Thus, the term-frequency vector $\mathcal{F}_1(d_m) \in \mathbb{R}^{7729}$ is represented as [1, 2, 1, 1, 1, 0,....,0], where only five (out of 7,729) elements have non-zero frequencies.

Unigrams					Bigrams				
we	manufacture	chicken	wing	thigh	we manufacture	manufacture chicken	chicken wing	wing chicken	chicken thigh
1	1	2	1	1	1	1	1	1	1
	✓	✓	✓	✓		✓			

Second, we determine which level-1 categories $c_{1j} \in \mathcal{C}_1$ the document d_m of the testing supplier belongs to most closely. This is done by computing the cosine similarity between the feature vector $\mathcal{F}_1(d_m) \in \mathbb{R}^{7729}$ of the testing supplier and the corresponding feature vector $x_{1j} \in \mathbb{R}^{7729}$ of each c_{1j}. For instance, the normalized cosine similarity for "Manufacturing" is 0.53, while it is 0.28 for"Wholesale and retail trade," 0.19 for "Agriculture, forestry and fishing," and zero for the other level-1

categories.

Bottom-Up Component

First, we extract the unigrams from d_m, and convert it into the relevant term-frequency vector $\mathcal{F}_6(d_m) \in \mathbb{R}^{6123}$. For example, suppose d_m = *"charalambides christis edam cheese charalambides butter"*, which contains five unique unigrams. If only two of these unigrams, i.e., "cheese" and "butter," overlapped with level-6 features, then $\mathcal{F}_6(d_m)$ = [1, 1, 0, 0, 0, ..., 0], wherein only two elements have non-zero frequencies out of a total of 6,123 dimensions.

charalambides	christis	edam	cheese	butter
2	1	1	1	1
			✓	✓

Second, we compare the cosine similarity between the testing supplier $\mathcal{F}_6(d_m) \in \mathbb{R}^{6123}$ with each level-6 category c_{6j}'s corresponding feature vector $x_{6j} \in \mathbb{R}^{6123}$. For example, "Butter oil" has a normalized cosine similarity of 0.0021, "Butter milk" has a normalized cosine similarity of 0.0011, "Lime growing" has a cosine similarity of 0, and so on.

Sandwich-Connection Component

The sandwich-connection component aims to identify the set of most likely SIC pathways (from level-1 category to level-6 category) that a testing supplier belongs to.

Among all 15,574 pathways $c_{1j} \rightarrow c_{6\ell}$, we can isolate the most likely pathways by examining the *product* of the normalized cosine similarity scores, shown in Table 4.10. As the last step, we select the pathways that are in the $q \in (0, 1)$ percentile based on their pathway similarity. Ultimately, our model predicts a supplier m into multiple pathways through level-1 to level-6 categories, i.e., $\mathcal{C}_m^* \triangleq \{\mathcal{C}_{1,m}^*, \mathcal{C}_{2,m}^*, \mathcal{C}_{3,m}^*, \mathcal{C}_{4,m}^*, \mathcal{C}_{5,m}^*, \mathcal{C}_{6,m}^*\}$

Table 4.10: Example of Sandwich-Connection Component for Predictions

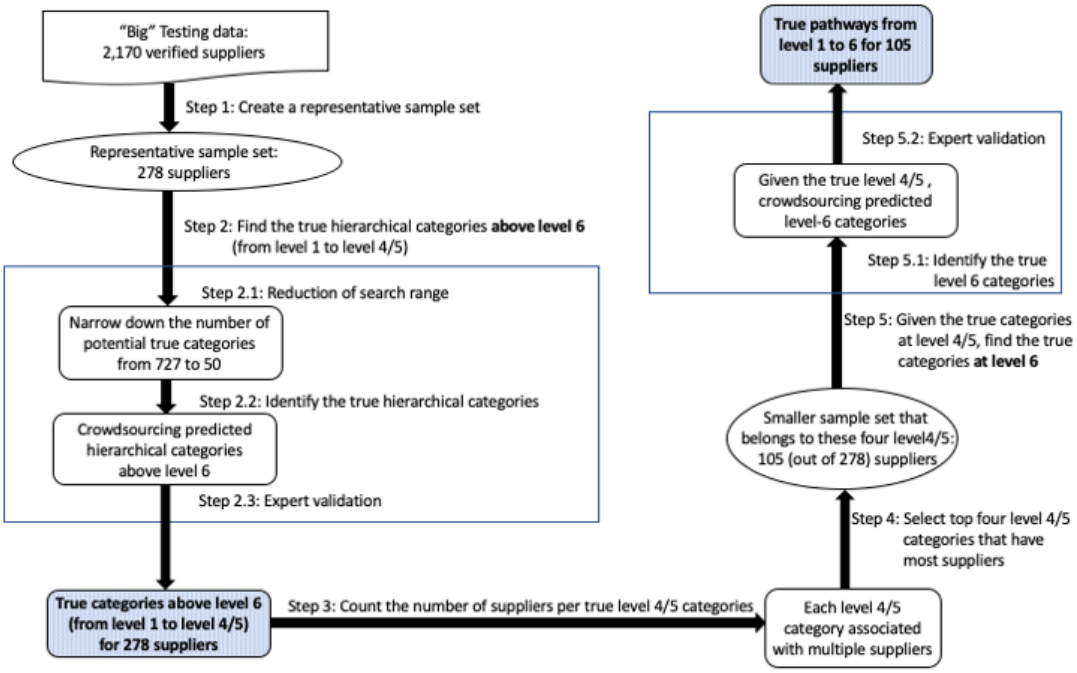
	Level 1 predicted	Level 1 normalized cosine similarity	Level 2-5 connected	Level 6 predicted	Level 6 normalized cosine similarity	Pathway Similarity
1	Manufacturing	0.53	...	Butter oil	0.0021	0.53×0.0021
2	Manufacturing	0.53	...	Butter milk	0.0011	0.53×0.0011
3	Manufacturing	0.53	...	Butter production	0.0010	0.53×0.0010
...	Manufacturing	0.53	...	...	...	...
9,189	Manufacturing	0.53	...	Roof lights made of plastic	0	0
9,190	Wholesale and retail trade	0.28	...	Butter	0.0020	0.28×0.0020
9,191	Wholesale and retail trade	0.28	...	Cheese	0.0019	0.28×0.0019
9,192	Wholesale and retail trade	0.28	...	Milking machines	0.0001	0.28×0.0001
...	Wholesale and retail trade	0.28	...	...	...	...
10,943	Wholesale and retail trade	0.28	...	Cinema kiosk	0	0
10,944	Agriculture, forestry and fishing	0.19	...	Butter	0.0020	0.19×0.0020
10,945	Agriculture, forestry and fishing	0.19	...	Lemon growing	0	0
...	Agriculture, forestry and fishing	0.19	...	...	...	...
11,456	Agriculture, forestry and fishing	0.19	...	Bean growing	0	0
11,457	...	0	...	...	...	0
...	...	0	...	...	...	0
15,574	...	0	...	...	...	0

4.9.2 Finding true labels

Prior research (e.g., Bragg et al. 2013, Budak et al. 2016) has emphasized the effectiveness of crowdsourcing in obtaining high-quality labels. In this study, we leveraged crowdsourcing via Amazon Mechanical Turk (MTurk) to construct true labels (i.e., true pathways) for a small sample of suppliers. To enhance the accuracy and reliability of the labels, we further incorporated a complementary validation checks by two procurement experts.

We begin with an overview of the steps of our true label gathering that combines crowdsourcing and expert validation in Figure 4.8. Step 1 draws a sample set of testing suppliers of Cranswick plc. In Step 2, we estimate the true hierarchical categories for the supplier sample for levels 4/5. This step involves (a) the design of simple experiments, (b) the aggregation of responses taking advantage of the wisdom of crowds, and (c) expert validation. In Step 3, the number of suppliers per category is counted, and in Step 4, the top four level 4/5 categories that include the maximum number of suppliers are selected to create a smaller sample set for identifying the true level 6. Finally, in Step 5, we repeat Step 2 to estimate the true level-6 categories for the smaller sample of suppliers. Below we describe each of these steps in detail.

Figure 4.8: Process of true label gathering



Step 1: Create a representative sample set. A sample of 278 suppliers was drawn from a total of 2,170 suppliers using a two-stage sampling approach. The first stage involved the selection of 68 suppliers who contributed to 80% of the total invoice value, while the second stage involved the random sampling of 10% (i.e., 210) of the remaining 2,102 suppliers. The resulting sample size of 278 suppliers represents 12.8% of the total supplier population.

Step 2: Find the true hierarchical categories for levels 4/5. Table 4.3 shows that before the expansion of SIC to level 6, there are 727 unique level 4/5 categories, with 191 at level 5 and 536 at level 4. For each of the 278 sampled suppliers, we asked MTurkers to find the most likely SIC categories from level 1 to level 4/5 the supplier belongs to, and asked two procurement experts to validate the provisional true labels. To ensure reliable predictions from level 1 to level 4/5, we segmented the tasks into three sub-steps.

Step 2.1: Reduction of Search Range. First, we narrowed down the number of potential level 4/5 categories for each supplier from 727 to a more manageable number. For each sampled supplier, we grouped the 727 level 4/5 categories into 73 groups based on the combined similarity scores from the three-component model. These groups were constructed such that each group contained 10 categories, with group 1 comprising the most similar ten level 4/5 categories (1st to 10th) and group 73 containing the least similar level 4/5 categories (721st to 727th). Figure 4.9 shows an example of Mturk task with the first two groups for a focal supplier.

To ascertain the likelihood of a supplier belonging to at least one of the categories within each group, we allocated five different MTurkers to check per supplier and each MTurker was asked 73 randomly ordered TRUE/FALSE questions. We launched 1,390 MTurk tasks (5 MTurkers per supplier $\times$ 278 suppliers, \$1.5 per task) and collected 101,470 TRUE/FALSE answers (73 answers per task $\times$ 1,390 tasks). To ensure no potentially relevant pathway was unintentionally removed, we kept the groups for which at least two MTurkers agreed TRUE. Table 4.11 shows that the agreed (at least 2 out of 5) "TRUE" answers fall within group 1 through group 5 among 278 sampled suppliers. The preliminary task suggested that MTurkers can be asked to search for true level 4/5 categories among 50 (not 727) without compromising accuracy.

Step 2.2: Identifying the true hierarchical categories. Based on 50 level 4/5 categories per supplier, we requested MTurkers to determine whether a focal supplier belonged to each category by responding to 50 TRUE/FALSE questions. Figure 4.10 shows an example of such MTurk task. To provide MTurkers with relevant information for their assessments, we provided the supplier's official website URL (269 out of 278 have official website URLs and website text data) and a snippet of purchase order records.

We ensured a sufficient number of responses per supplier by allocating five dif-

Figure 4.9: Reduction of Search Range MTurk Example (Step 2.1)

Level1	Level 4/5	Group_number
AGRICULTURE, FORESTRY AND FISHING	Growing of vegetables and melons, roots and tubers	1
AGRICULTURE, FORESTRY AND FISHING	Mixed farming	1
MANUFACTURING	Other processing and preserving of fruit and vegetables	1
AGRICULTURE, FORESTRY AND FISHING	Gathering of wild growing non-wood products	1
AGRICULTURE, FORESTRY AND FISHING	Growing of other tree and bush fruits and nuts	1
AGRICULTURE, FORESTRY AND FISHING	Growing of spices, aromatic, drug and pharmaceutical crops	1
WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES	Wholesale of fruit and vegetables	1
MANUFACTURING	Processing and preserving of potatoes	1
MANUFACTURING	Manufacture of sugar confectionery	1
WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES	Retail sale of fruit and vegetables in specialised stores	1
MANUFACTURING	Manufacture of agricultural and forestry machinery (other than agricultural tractors)	2
AGRICULTURE, FORESTRY AND FISHING	Plant propagation	2
AGRICULTURE, FORESTRY AND FISHING	Growing of other non-perennial crops	2
AGRICULTURE, FORESTRY AND FISHING	Post-harvest crop activities	2
MANUFACTURING	Manufacture of soft drinks; production of mineral waters and other bottled waters	2
MANUFACTURING	Manufacture of cider and other fruit wines	2
MANUFACTURING	Manufacture of prepared feeds for farm animals	2
MANUFACTURING	Tea processing	2
MANUFACTURING	Manufacture of condiments and seasonings	2
MANUFACTURING	Manufacture of wine from grape	2

Table 4.11: Reduction of Search Range MTurk Results (Step 2.1)

	Number of suppliers (Each tagged by 5 MTurkers)						
	0/5 TRUE	1/5 TRUE	2/5 TRUE	3/5 TRUE	4/5 TRUE	5/5 TRUE	Total
Group 1	9	11	44	68	60	86	278
Group 2	23	48	68	45	35	59	278
Group 3	51	81	43	53	19	31	278
Group 4	105	121	27	10	9	6	278
Group 5	125	114	32	3	4	0	278
Group 6	183	95	0	0	0	0	278
...	...	...	...	...	...	...	278
Group 73	278	0	0	0	0	0	278

ferent MTurkers to label each supplier, resulting in the completion of 1,390 MTurk tasks (5 MTurkers per supplier $\times$ 278 suppliers), with each task being compensated at a rate of \$2. The resulting dataset consisted of 69,500 TRUE/FALSE responses (50 responses per task $\times$ 1,390 tasks). Based on the majority of TRUE answers received from the MTurkers for each supplier (i.e., at least 3 out of 5), we identified the level 4/5 categories that served as the provisional true labels for each supplier.

Figure 4.10: Identifying True Level 4/5 Example (Step 2.2)

Level1	Level 4/5	Option index
WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES	Retail sale of antiques including antique books, in stores	1
WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES	Retail sale of fish, crustaceans and molluscs in specialised stores	2
MANUFACTURING	Manufacture of wire products, chain and springs	3
WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES	Retail sale of meat and meat products in specialised stores	4
WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES	Retail sale of electrical household appliances in specialised stores	5
WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES	Retail sale of fruit and vegetables in specialised stores	6
WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES	Retail sale of music and video recordings in specialised stores	7
WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES	Retail sale of flowers, plants, seeds, fertilisers, pet animals and pet food in specialised stores	8
WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES	Retail sale of beverages in specialised stores	9
WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES	Other retail sale of food in specialised stores	10
...	...	...
...	...	...
...	...	...
WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES	Non-specialised wholesale of food, beverages and tobacco	46
WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES	Retail sale via stalls and markets of other goods	47
MINING AND QUARRYING	Support activities for petroleum and natural gas extraction	48
WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES	Agents involved in the sale of furniture, household goods, hardware and ironmongery	49
WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES	Wholesale of petroleum and petroleum products	50

Step 2.3: Expert validation. To improve the reliability of the initial true labels gathered via crowdsourcing for the 278 suppliers, we engaged two procurement experts for verification. These experts undertook the same labeling task and reviewed the answers provided by the MTurkers. Through this verification process, we identified a total of 655 true level 4/5 categories, including 107 unique categories. Among these suppliers, 84 (30%), 99 (36%), 58 (21%), and 37 (13%) were respectively associated with one, two, three, and more than three true level 4/5 categories. Verifying the true labels for these suppliers was an exhaustive effort, requiring three months to complete. However, this rigorous process was critical in ensuring the accuracy of the true labels for our sample of suppliers.

Step 3: Count the number of suppliers for each true level 4/5 categories. The aforementioned Steps 1 and 2 obtained accurate hierarchical categories at level $i \in \{1, 2, 3, 4, 5\}$. The next three steps aim to identify the correct category at level 6. In our sample set of 278 suppliers, 107 (out of 727) unique categories were identified at level 4/5, which in turn encompassed 3,258 level 6 categories. Due to the substantial time and cost involved in obtaining true labels for a vast number of potential level 6 categories, we further narrowed down our focus to a smaller set

of suppliers. To do so, in Step 3, we counted the number of suppliers in each true level 4/5 category. The top 10 level 4/5 categories, along with their corresponding supplier count, are presented in Table 4.12.

Table 4.12: Level 4/5 true categories and associated number of suppliers (Step 3). Note that each supplier can belong to multiple level 4/5 categories.

	Level 4/5 true category for 278 suppliers	# of suppliers (total 278)	Selected
1	Production of meat and poultry meat products	47	✓
2	Processing and preserving of meat	30	✓
3	Other processing and preserving of fruit and vegetables	23	✓
4	Butter and cheese production	21	✓
5	Repair of machinery	19	×
6	Processing and preserving of poultry meat	18	×
7	Manufacture of plastic packing goods	16	×
8	Manufacture of prepared feeds for farm animals	17	×
9	Manufacture of paper and paperboard containers other than sacks and bags	14	×
10	Manufacture of other articles of paper and paperboard	14	×

Note: Each supplier can belong to multiple level 4/5 categories.

Step 4: Select top four 4/5 categories and create a smaller sample set. In Step 4, our objective was to restrict the sample size to a more manageable level to find true level 6, while still maintaining an adequate degree of statistical rigor. Based on these considerations, we selected a subset of four level 4/5 categories that comprised of more than 20 suppliers. These categories are as follows: "Production of meat and poultry meat products,""Processing and preserving of meat,""Other processing and preserving of fruit and vegetables" and "Butter and cheese production". The selected categories comprised a total of 121 suppliers, of which 105 were unique. Therefore, we reduced the number of suppliers in our sample set from 278 to 105. It is worth noting that each level 4/5 category comprises different numbers of level 6 categories, with a range from a minimum of 9 to a maximum of 179. Table 4.13 provides a summary of this information.

Step 5: Find the true level 6 categories conditioned on true level 4/5. In Step 5, we conducted an additional phase of validation using MTurk and expert

Table 4.13: Summary of Level 6 for the selected four level 4/5 categories (Step 4). Note that each supplier can belong to multiple level 4/5 categories.

Level 4/5 category	Number of suppliers	Number of level 6		
		Min	Max	Mean
Production of meat and poultry meat products	47	46	179	76
Processing and preserving of meat	30	62	179	92
Other processing and preserving of fruit and vegetables	23	52	151	69
Butter and cheese production	21	9	55	29
Total	105 (unique)	9	179	64

Note: Each supplier can belong to multiple level 4/5 categories.

assessments for 105 selected suppliers, focusing on pinpointing their accurate categories at level 6, conditional on their confirmed categories at levels 4/5. To facilitate this, we created a MTurk task that presented participants with 40 to 75 TRUE/-FALSE questions, each aimed at verifying whether a specific supplier was associated with a given level 6 category. Due to suppliers being associated with multiple categories at levels 4/5, the array of potential level 6 categories they could belong to varied. For instance, as illustrated in Figure 4.11, one particular supplier was linked to two level-4/5 categories ("Butter and cheese production" and "Liquid milk and cream production"), which together encompass 25 distinct level 6 categories.

To ensure that each MTurk task contained 40 to 75 questions, we grouped two or more suppliers into one task if the supplier had less than 40 potential level 6 categories, or split a supplier into two or three tasks if the supplier had more than 75 potential level 6 categories. The MTurk task allocation process is detailed in Table 4.14, showing that 105 suppliers comprised of 122 MTurk tasks.

A total of 610 MTurk tasks were completed, with five different MTurkers assigned to label each task, and each task being compensated at a rate of $2. Based on the majority of TRUE answers received from the MTurkers for each supplier (i.e., at least 3 out of 5), we further asked two procurement experts to validate the responses. As a result, a total of 416 true level 6 categories were identified for 105 suppliers.

Figure 4.11: Identifying True Level 6 Example (Step 5)

Level 4/5	Level 6	Option index
Butter and cheese production	Butter blending	1
Butter and cheese production	Butter milk	2
Butter and cheese production	Butter oil	3
Butter and cheese production	Butter production	4
Butter and cheese production	Butterfat	5
Butter and cheese production	Cheese	6
Butter and cheese production	Curd production	7
Butter and cheese production	Dairy preparation of cheese and butter	8
Butter and cheese production	Processed cheese	9
Liquid milk and cream production	Clotted cream	10
Liquid milk and cream production	Cream (sterilised)	11
Liquid milk and cream production	Cream from fresh homogenized liquid milk	12
Liquid milk and cream production	Cream production	13
Liquid milk and cream production	Double cream	14
Liquid milk and cream production	Heat treatment of milk	15
Liquid milk and cream production	Homogenised milk production	16
Liquid milk and cream production	Milk sterilising	17
...	...	...
...	...	...
Liquid milk and cream production	Sterilised cream	25

Table 4.14: MTurk tasks allocation for level-6 true labels

	Bundle of four suppliers (each with < 20 level 6)	Bundle of two suppliers (each with 20-39 level 6)	1 task (each with 40-75 level 6)	Split into 2 tasks (each with 76-120 level 6)	Split into 3 tasks (each with > 120 level 6)	Total
# of suppliers	8	4	68	19	6	105
# of MTurk tasks	2	2	68	38	12	122

4.9.3 Detailed Explanation for Benchmark Models

Top-down Model

The benchmark top-down model predicts supplier m level-by-level along the SIC pathway until reaching the level 6. At any level $i \in \{1, 2, 3, 4, 5, 6\}$, we use the selectively parameter $q \in (0, 1)$ to choose predicted categories c_{ij} whose cosine similarities are in the q^{th} percentile.

It starts with level 1 prediction by calculating the cosine similarities between d_m and each $c_{1j} \in \mathcal{C}_1$ ($|\mathcal{C}_1| = 21$). Any c_{1j} will be kept if its normalized cosine similarity is in the $q \in (0, 1)$ percentile. Table 4.15 shows an example. Say $q = 0.99$ gives us the cut-off value as 0.311, then only one level-1 category "Manufacturing" is kept. Consequently, the top-down model narrows the choices at level 2, only considering the subset of $\mathcal{C}_2$ that belongs to survived c_{1j}.

Table 4.15: Benchmark Top-down Model for Level 1 Predictions (Example with a Single Supplier)

	Level 1 predicted	Level 1 normalized cosine similarity	Keep
1	Manufacturing	0.312	✓
2	Agriculture, forestry and fishing	0.307	×
3	Professional, scientific and technical activities	0.143	×
4	Transportation and storage	0.085	×
...	...	...	×
...	...	...	×
20	Construction	0.000	×
21	Wholesale and retail trade	0.000	×

In the level 2 prediction, we calculate the cosine similarities between d_m and each c_{2j} in the survived subset, and use the same selectively parameter $q \in (0, 1)$ to choose predicted categories c_{2j} whose cosine similarities are in the q^{th} percentile. Continuing with the example above, we show the similarities between d_m and 24 c_{2j} within "Manufacturing" in Table 4.16. Applying $q = 0.99$ leads to two categories "Manufacture of food products" and "Manufacture of basic metals" survived. Consequently, the top-down model narrows the choices at level 3 within these two survived level-2 categories. We repeat the prediction down to level 6.

Table 4.16: Benchmark Top-down Model for Level 2 Predictions (Example with a Single Supplier)

	Survival Level 1	Level 2 predicted	Level 2 normalized cosine similarity	Keep
1	Manufacturing	Manufacture of food products	0.367	✓
2	Manufacturing	Manufacture of basic metals	0.327	✓
3	Manufacturing	Manufacture of textiles	0.102	×
...	...	...	...	×
23	Manufacturing	Manufacture of leather and related products	0.000	×
24	Manufacturing	Manufacture of tobacco products	0.000	×

Bottom-Up Model

The benchmark bottom-up model is similar to the bottom-up component. It starts by calculating the cosine similarities between d_m and each $c_{6j} \in \mathcal{C}_6$ ($|\mathcal{C}_6| = 15574$). For each c_{6j}, we can trace the corresponding pathway up to level 1, and directly

use the cosine similarity at level 6 to represent pathway similarity (See Table 4.17). Then, we apply the selectively parameter $q \in (0, 1)$ to choose the pathways that have pathway similarities in the q^{th} percentile.

Table 4.17: Benchmark Bottom-up Model Predictions (Example with a Single Supplier)

	Level 6 predicted	Level 6 normalized cosine similarity	Level 1-5 traced	Pathway Similarity
1	Butter oil	0.0023	...	0.0023
2	Cocoa butter	0.0021	...	0.0021
3	Shea butter	0.0019	...	0.0019
4	Peanut butter	0.0018	...	0.0018
...	...	...	...	
...	...	...	...	
15573	Pork pie	0.0000	...	0.0000
15574	Lime Growing	0.0000	...	0.0000

4.9.4 F1 Scores for All Combinations of Text

The following tables detail the average F1 scores for different text combinations utilized within the three-component model, as well as the benchmark top-down and bottom-up models. Columns that are part of the combinations previously presented in Table 4.5 are shaded in yellow for emphasis.

Table 4.18: Average F1 Score in Three-Component Model with All Combinations of Text. Here, F1 at levels 1-5 are aggregated with 278 suppliers in the set $\mathcal{M}$, and level 6 is aggregated with 105 suppliers in the set $\mathcal{M}'$. In all models, $q = \hat{q} = 0.99$.

	General description	Specific descriptions			All texts	Best data source for each entry
	d_m^{gen} (1)	d_m^{spe} (2)	d_m^{PO} (3)	$d_m^{spe} \cup d_m^{PO}$ (4)	d_m (5)	flex (6)
Level 1	0.898	0.849	0.723	0.848	0.858	0.935
Level 2	0.575	0.600	0.467	0.563	0.569	0.675
Level 3	0.437	0.464	0.366	0.445	0.449	0.530
Level 4	0.385	0.412	0.297	0.376	0.382	0.457
Level 5	0.379	0.407	0.282	0.370	0.377	0.449
Level 6	0.349	0.368	0.179	0.359	0.367	0.424

Table 4.19: Average F1 Score in Benchmark Top-down Model with All Combinations of Text. Here, F1 at levels 1-5 are aggregated with 278 suppliers in the set $\mathcal{M}$, and level 6 is aggregated with 105 suppliers in the set $\mathcal{M}'$. In all models, $q = \hat{q} = 0.99$.

	General description	Specific descriptions			All texts
	d_m^{gen} (1)	d_m^{spe} (2)	d_m^{PO} (3)	$d_m^{spe} \cup d_m^{PO}$ (4)	d_m (5)
Level 1	0.849	0.746	0.610	0.742	0.802
Level 2	0.579	0.544	0.385	0.510	0.561
Level 3	0.387	0.407	0.288	0.414	0.438
Level 4	0.294	0.327	0.231	0.333	0.356
Level 5	0.277	0.312	0.221	0.319	0.342
Level 6	0.158	0.190	0.138	0.189	0.204

Table 4.20: Average F1 Score in Benchmark Bottom-up Model with All Combinations of Text. Here, F1 at levels 1-5 are aggregated with 278 suppliers in the set $\mathcal{M}$, and level 6 is aggregated with 105 suppliers in the set $\mathcal{M}'$. In all models, $q = \hat{q} = 0.99$.

	General description	Specific descriptions			All texts
	d_m^{gen} (1)	d_m^{spe} (2)	d_m^{PO} (3)	$d_m^{spe} \cup d_m^{PO}$ (4)	d_m (5)
Level 1	0.563	0.638	0.619	0.663	0.642
Level 2	0.372	0.421	0.320	0.421	0.403
Level 3	0.289	0.342	0.239	0.342	0.321
Level 4	0.258	0.306	0.200	0.306	0.283
Level 5	0.254	0.305	0.200	0.305	0.281
Level 6	0.242	0.246	0.274	0.301	0.230

Bibliography

Abadie A, Diamond A, Hainmueller J (2010) Synthetic control methods for comparative case studies: Estimating the effect of Californias tobacco control program. *Journal of the American Statistical Association* 105(490):493–505.

Abadie A, Diamond A, Hainmueller J (2011) Synth: An R package for synthetic control methods in comparative case studies. *Journal of Statistical Software* 42(13), URL `https://www.jstatsoft.org/issue/view/v042`.

Abadie A, Diamond A, Hainmueller J (2015) Comparative politics and the synthetic control method. *American Journal of Political Science* 59(2):495–510.

Adalja A, Liaukonytė J, Wang E, Zhu X (2023) Gmo and non-GMO labeling effects: Evidence from a quasi-natural experiment. *Marketing Science* 42(2):233–250.

Akanksh AM, Nayak MS, Nishith S, Pandit SN, Sunkad S, Deenadhayalan P, Gangadhara S (2023) Automated invoice data extraction using image processing. *IAES International Journal of Artificial Intelligence* 12(2):514.

Almenberg J, Dreber A (2011) When does the price affect the taste? Results from a wine experiment. *Journal of Wine Economics* 6(1):111–121.

Arkhangelsky D, Athey S, Hirshberg DA, Imbens GW, Wager S (2021) Synthetic difference-in-differences. *American Economic Review* 111(12):4088–4118.

Asch SE (1955) Opinions and social pressure. *Scientific American* 193(5):31–35.

Ashenfelter O, Jones GV (2013) The demand for expert opinion: Bordeaux wine. *Journal of Wine Economics* 8(3):285–293.

Beil DR, Chen Q, Duenyas I, See BD (2018) When to deploy test auctions in sourcing.

Manufacturing & Service Operations Management 20(2):232–248.

Beil DR, Wein LM (2003) An inverse-optimization-based auction mechanism to support a multiattribute RFQ process. *Management Science* 49(11):1529–1545.

Berger J, Humphreys A, Ludwig S, Moe WW, Netzer O, Schweidel DA (2020) Uniting the tribes: Using text for marketing insight. *Journal of Marketing* 84(1):1–25.

Berman R, Israeli A (2022) The value of descriptive analytics: Evidence from online retailers. *Marketing Science* 41(6):1074–1096.

Bondi T, Rossi M, Stevens R (2023) The good, the bad and the picky: Reference dependence and the reversal of product ratings. *Proceedings of the 24th ACM Conference on Economics and Computation*, 297–297.

Bragg J, Weld D, et al. (2013) Crowdsourcing multi-label classification for taxonomy creation. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 1, 25–33.

Budak C, Goel S, Rao JM (2016) Fair and balanced? Quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly* 80(S1):250–271.

Büschken J, Allenby GM (2016) Sentence-based text analysis for customer reviews. *Marketing Science* 35(6):953–975.

Cai L, Hofmann T (2004) Hierarchical document categorization with support vector machines. *Proceedings of the 13th ACM International Conference on Information and Knowledge Management*, 78–87.

Ceci M, Malerba D (2007) Classifying web documents in a hierarchy of categories: a comprehensive study. *Journal of Intelligent Information Systems* 28(1):37–78.

Cerri R, Barros RC, De Carvalho AC (2014) Hierarchical multi-label classification using local neural networks. *Journal of Computer and System Sciences* 80(1):39–56.

Cesa-Bianchi N, Gentile C, Zaniboni L (2006) Hierarchical classification: combining bayes with svm. *Proceedings of the 23rd International Conference on Machine Learning*, 177–184.

Cevahir A, Murakami K (2016) Large-scale multi-class and hierarchical product categoriza-

tion for an e-commerce giant. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 525–535.

Chae BK (2015) Insights from hashtag supplychain and twitter analytics: Considering twitter and twitter data for supply chain practice and research. *International Journal of Production Economics* 165:247–259.

Chang J, Gerrish S, Wang C, Boyd-Graber JL, Blei DM (2009) Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems*, 288–296.

Chaturvedi A, Beil DR, Martínez-de Albéniz V (2014) Split-award auctions for supplier retention. *Management Science* 60(7):1719–1737.

Chen H, Chiang RH, Storey VC (2012a) Business intelligence and analytics: From big data to big impact. *MIS quarterly* 1165–1188.

Chen Y, Liu Y, Zhang J (2012b) When do third-party product reviews affect firm value and what can firms do? The case of media critics and professional movie reviews. *Journal of Marketing* 76(2):116–134.

Chen Y, Xie J (2005) Third-party product review and firm marketing strategy. *Marketing Science* 24(2):218–240.

Copas JB, Li H (1997) Inference for non-random samples. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 59(1):55–95.

Costa E, Lorena A, Carvalho A, Freitas A (2007) A review of performance evaluation measures for hierarchical classifiers. *Evaluation Methods for Machine Learning II: Papers from the AAAI-2007 Workshop*, 1–6.

Cranswick (2023) Cranswick plc annual report accounts. URL `https://s3.eu-west-1.amazonaws.com/cranswick-2021/Interim-statement-FY23.pdf`.

Davenport TH (2013) Analytics 3.0. *Harvard Business Review.* URL `https://hbr.org/2013/12/analytics-30`.

Davenport TH, Patil D (2022) Is data scientist still the sexiest job of the 21st century?

Harvard Business Review. URL `https://hbr.org/2022/07/is-data-scientist-still-the-sexiest-job-of-the-21st-century`.

Diehl K, Poynor C (2010) Great expectations?! Assortment size, expectations, and satisfaction. *Journal of Marketing Research* 47(2):312–322.

Dittrich J, Julka R, Mercker BU, Riedstra P (2020) The role of spend analytics in the next normal. *McKinsey Insights.* URL `https://www.mckinsey.com/business-functions/operations/our-insights/the-role-of-spend-analytics-in-the-next-normal`.

Domingos P (2012) A few useful things to know about machine learning. *Communications of the ACM.* 55(10):78–87.

Duenyas I, Hu B, Beil DR (2013) Simple auctions for supply contracts. *Management Science* 59(10):2332–2342.

Dumais S, Chen H (2000) Hierarchical classification of web content. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 256–263.

Elmaghraby WJ (2000) Supply contract competition and sourcing policies. *Manufacturing & Service Operations Management* 2(4):350–371.

Esuli A, Fagni T, Sebastiani F (2008) Boosting multi-label hierarchical text categorization. *Information Retrieval* 11(4):287–313.

Farronato C, Zervas G (2022) Consumer reviews and regulation: Evidence from NYC restaurants. Technical report, National Bureau of Economic Research.

Fine CH, Porteus EL (1989) Dynamic process improvement. *Operations Research* 37(4):580–591.

Fogarty JJ (2012) Expert opinion and cuisine reputation in the market for restaurant meals. *Applied Economics* 44(31):4115–4123.

Friberg R, Grönqvist E (2012) Do expert reviews affect the demand for wine? *American Economic Journal: Applied Economics* 4(1):193–211.

Garcia LV (2021) BCG: Procurement of the future. *Boston Consulting Group, Procurement*

URL https://procurementmag.com/digital-procurement/bcg-procurement-future.

George G, Haas MR, Pentland A (2014) Big data and management. *Academy of Management Journal* 57(2):321–326.

Gergaud O, Storchmann K, Verardi V (2015) Expert opinion and product quality: Evidence from New York City restaurants. *Economic Inquiry* 53(2):812–835.

Gerstner E (1985) Do higher prices signal higher quality? *Journal of Marketing Research* 22(2):209–215.

Ginsburgh V (2003) Awards, success and aesthetic quality in the arts. *Journal of Economic Perspectives* 17(2):99–111.

Griffiths TL, Steyvers M (2004) Finding scientific topics. *Proceedings of the National Academy of Sciences* 101(suppl 1):5228–5235.

Guo J, Che W, Wang H, Liu T (2014) Revisiting embedding features for simple semi-supervised learning. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 110–120.

Hackmann MB, Kolstad JT, Kowalski AE (2015) Adverse selection and an individual mandate: When theory meets practice. *American Economic Review* 105(3):1030–66.

Hasenbein JJ, Gray P, Greenberg HJ (2010) Risk and optimization in an uncertain world. *INFORMS TutORials in Operations Research* .

Hayward T (2021) Whats wrong with the Michelin guide? *Finantial Times.* URL https://www.ft.com/content/e622ec53-ea9f-487a-a434-747f13f5ffa0.

Hilger J, Rafert G, Villas-Boas S (2011) Expert opinion and the demand for experience goods: An experimental approach in the retail wine market. *Review of Economics and Statistics* 93(4):1289–1296.

Holden N, Freitas AA (2006) Hierarchical classification of g-protein-coupled receptors with a pso/aco algorithm. *Proceedings of the IEEE Swarm Intelligence Symposium*, 77–84.

Hollenbeck B (2018) Online reputation mechanisms and the decreasing value of chain affiliation. *Journal of Marketing Research* 55(5):636–654.

Johnson C, Surlemont B, Nicod P, Revaz F (2005) Behind the stars: a concise typology of michelin restaurants in Europe. *Cornell Hotel and Restaurant Administration Quarterly* 46(2):170–187.

Kovács B, Sharkey AJ (2014) The paradox of publicity: How awards can negatively affect the evaluation of quality. *Administrative Science Quarterly* 59(1):1–33.

Kraljic P (1983) Purchasing must become supply management. *Harvard Business Review* 61(5):109–117.

Krosnick JA (1999) Survey research. *Annual Review of Psychology* 50(1):537–567.

Laffont JJ, Tirole J (1993) *A theory of incentives in procurement and regulation* (MIT press).

Li C, Wan Z (2017) Supplier competition and cost improvement. *Management Science* 63(8):2460–2477.

Li KT (2020) Statistical inference for average treatment effects estimated by synthetic control methods. *Journal of the American Statistical Association* 115(532):2068–2083.

Li X, Hitt LM (2008) Self-selection and information role of online product reviews. *Information Systems Research* 19(4):456–474.

Maas A, Daly RE, Pham PT, Huang D, Ng AY, Potts C (2011) Learning word vectors for sentiment analysis. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 142–150.

Mao Y, Tian J, Han J, Ren X (2019) Hierarchical text classification with reinforced label assignment. *Proceeding of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, volume 1.

McCallum A, Rosenfeld R, Mitchell TM, Ng AY (1998) Improving text classification by shrinkage in a hierarchy of classes. *In Proceedings of 15th International Conference on Machine Learning*, volume 98, 359–367.

McKinsey (2021) Actionable spend insights with orpheus. URL `https://www.mckinsey.`

com/business-functions/operations/how-we-help-clients/product-development-procurement/actionable-spend-insights-orpheus.

Mišić VV, Perakis G (2020) Data analytics in operations management: A review. *Manufacturing & Service Operations Management* 22(1):158–169.

Mittal V, Han K, Lee JY, Sridhar S (2021) Improving business-to-business customer satisfaction programs: Assessment of asymmetry, heterogeneity, and financial impact. *Journal of Marketing Research* 58(4):615–643.

Moe WW, Trusov M (2011) The value of social dynamics in online product ratings forums. *Journal of Marketing Research* 48(3):444–456.

Morgan NA, Rego LL (2006) The value of different customer satisfaction and loyalty metrics in predicting business performance. *Marketing Science* 25(5):426–439.

Olsen TL, Tomlin B (2020) Industry 4.0: Opportunities and challenges for operations management. *Manufacturing & Service Operations Management* 22(1):113–122.

Ospina D (2018) How the best restaurants in the world balance innovation and consistency. *Harvard Business Review.* URL https://hbr.org/2018/01/how-the-best-restaurants-in-the-world-balance-innovation-and-consistency.

Peng H, Li J, He Y, Liu Y, Bao M, Wang L, Song Y, Yang Q (2018) Large-scale hierarchical text classification with recursively regularized deep graph-cnn. *Proceedings of the 2018 world wide web conference*, 1063–1072.

Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543, URL http://www.aclweb.org/anthology/D14-1162.

Puranam D, Narayan V, Kadiyali V (2017) The effect of calorie posting regulation on consumer opinion: a flexible latent dirichlet allocation model with informative priors. *Marketing Science* 36(5):726–746.

Ramos J, et al. (2003) Using TF-IDF to determine word relevance in document queries. *Proceedings of the 1st Instructional Conference on Machine Learning*, volume 242, 29–48.

Rao AR, Monroe KB (1989) The effect of price, brand name, and store name on buyers perceptions of product quality: An integrative review. *Journal of Marketing Research* 26(3):351–357.

Röder M, Both A, Hinneburg A (2015) Exploring the space of topic coherence measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399–408.

Rossi M (2021) Quality disclosures and disappointment: Evidence from the Academy Awards. *Proceedings of the 22nd ACM Conference on Economics and Computation*, 790–791.

Sands D (2020) The role of third parties in value creation and capture: Why Michelin stars may not be a good thing. *Academy of Management Proceedings*, 21105, number 1.

Shen D, Ruvini JD, Sarwar B (2012) Large-scale item categorization for e-commerce. *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, 595–604.

Shin M, Shin J, Ghili S, Kim J (2023) The impact of the gig economy on product quality through the labor market: Evidence from ridesharing and restaurant quality. *Management Science* 69(5):2620–2638.

Silla CN, Freitas AA (2011) A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery* 22(1):31–72.

Statista (2017) How much do movie critic reviews influence your decision to see a movie? `https://www.statista.com/statistics/682930/movie-critic-reviews-influence/`, [Online; accessed 08-September-2020].

Tarawneh AS, Hassanat AB, Chetverikov D, Lendak I, Verma C (2019) Invoice classification using deep features and machine learning techniques. *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology*, 855–859.

Tirunillai S, Tellis GJ (2014) Mining marketing meaning from online chatter: Strategic brand analysis of big data using Latent Dirichlet Allocation. *Journal of Marketing*

Research 51(4):463–479.

UK Office for National Statistics (2007) Standard Industrial Classification. URL https://www.ons.gov.uk/methodology/classificationsandstandards/ukstandardindustrialclassificationofeconomicactivities/uksic2007.

US Census Bureau (2022) Statistics for industry groups and industries. annual survey of manufactures: 2020. URL https://www.census.gov/library/publications/2020/econ/e20-asm.html.

Vickrey W (1961) Counterspeculation, auctions, and competitive sealed tenders. *Journal of Finance* 16(1):8–37.

Wan Z, Beil DR (2009) RFQ auctions with supplier qualification screening. *Operations Research* 57(4):934–949.

Webb J (2017) What is the kraljic matrix? *Forbes.* URL https://www.forbes.com/sites/jwebb/2017/02/28/what-is-the-kraljic-matrix/?sh=24f54588675f.

Winer RS (1986) A reference price model of brand choice for frequently purchased products. *Journal of Consumer Research* 13(2):250–256.

Xu Y (2017) Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis* 25(1):57–76.

Zhang M, Luo L (2023) Can consumer-posted photos serve as a leading indicator of restaurant survival? evidence from yelp. *Management Science* 69(1):25–50.

Zhao H, Du L, Buntine W, Liu G (2017) MetaLDA: A topic model that efficiently incorporates meta information. *IEEE International Conference on Data Mining*, 635–644.