Discovering the value of unstructured data in business settings

Nan Yang

100291479

Norwich Business School

This dissertation is submitted for the degree of

Doctor of Philosophy

September 2023

ABSTRACT

With the increasing amount of unstructured data in business settings, the analysis of unstructured data is reshaping business practices in many industries. The implementation of unstructured data analysis will eventually have dominant presence in all department of organisations thus contributing to the organisations. This dissertation focuses the most widely utilised unstructured data-textual data within the organisation. A variety of techniques has been applied in three studies to discover the information within the unstructured textual data. Study I proposed a dynamic model that incorporates values from topic membership, an outcome variable from Latent Dirichlet Allocation (a probabilistic topic model), with sentiment analysis for rating prediction. A variety of machine learning algorithms are employed to validate the model. Study II focused on the exploration of online reviews from customers in the OFD domain. In addition, this study examines the outcomes of franchising in the service sector from the customer's perspective. This study identifies key issues during the processes of producing and delivering product/services from service providers to customers in service industries using a large-scale dataset. Study III extends the data scope to the firm-level data. Latent signals are discovered from companies' self-descriptions. In addition, the association between the signals and the organisation context of the entrepreneurship is also examined, which could display the heterogeneity of various signals across different organisation context.

Publications

Yang, N., Korfiatis, N., Zissis, D. and Spanaki, K. (2023) Incorporating topic membership in review rating prediction from unstructured data: a gradient boosting approach. *Annals of Operations Research*. doi:10.1007/s10479-023-05336-z.

Yang, N., Dousios, D., Korfiatis, N. and Chalvatzis, K. (2024). Mapping the signaling environment between sustainability‐focused entrepreneurship and investment inputs: A topic modeling approach. *Business Strategy and the Environment.* doi:10.1002/bse.3748.

Table of Contents

Discovering the value of unstructured data in business settings I

ABSTRACT II

Publications III

List of tables VIII

List of Figures X

Acknowledgement XII

CHAPTER 1 Introduction 1

1.1 Research background 1

1.2 Research questions 3

1.3 Research methodology 5

1.4 Research outline 13

CHAPTER 2 Literature review 15

2.1 Unstructured data 15

2.1.1 Unstructured data and structured data 15

2.1.2 Types of unstructured data 17

2.1.3 Characteristic of unstructured data 19

2.1.4 Unstructured data analytics 21

2.2 Textual data 23

2.2.1 Definitive characteristics of text 23

2.2.2 Features of text at different levels 25

2.3 Content analysis VS Text analytics 28

2.3.1 Content analysis definition 28

2.3.2 From content analysis to text mining 30

2.3.3 Syntax versus semantics 33

2.4 Research gaps 34

CHAPTER 3 Incorporating topic membership in rating prediction from unstructured data: a gradient boosting approach 36

3.1 Introduction 36

3.2 Literature review 39

3.2.1 Sentiment analysis 39

3.2.2 Topic models on user-generated content 42
3.2.3 Rating prediction 43
3.2.4 Research gaps 47

3.3 Data and methods 48

3.3.1 Dataset 49
3.3.2 Sentiment analysis 51
3.3.3 Latent Dirichlet Allocation (LDA) 52
3.3.4 Supervised machine learning techniques in the classification 55
3.3.5 Model performance measurement metrics 61
3.3.6 Feature selection 64

3.4 Results 65

3.4.1 Baseline sentiment calculation 65
3.4.2 Corpus pre-processing for topic modelling 65
3.4.3 LDA model estimation and hyperparameter tuning 66
3.4.4 Topic identification 68
3.4.5 Classification using 6 algorithms 70
3.4.6 Incorporating topic membership in sentiment text detection 75
3.4.7 Rating score prediction 81
3.4.8 Shapley Additive explanations (SHAP) for each feature 82

3.5 Discussion and implications 84

3.5.1 Discussion 84
3.5.2 Theoretical implications 86
3.5.3 Practical implications 87

3.6 Conclusions, limitations, and future research 88

CHAPTER 4 Identifying latent service dimensions in service supply chains: an analysis of franchised and independent businesses 90

4.1 Introduction 90

4.2 Theoretical background 93

4.2.1 Service supply chain management in the OFD sector 94
4.2.2 The importance of online reviews 96
4.2.3 Franchising as a strategy 98
4.2.4 Research gaps 100

4.3 Data and methods 101

4.3.1 Data 101
4.3.2 Research design 103
4.3.3 Text pre-processing 103

4.3.4 Collocation analysis techniques 106
4.3.5 Structural topic model 110
4.3.6 Model estimation 113

4.4 Results 114
4.4.1 Text statistics 114
4.4.2 Collocation analysis 116
4.4.3 Topic solution 122
4.4.4 Comparison between franchised and independent restaurants 126
4.4.5 The interaction of the restaurant type 129
4.4.6 Digging into franchised restaurants 129

4.5 Discussion and implications 133
4.5.1 Theoretical implications 134
4.5.2 Implications for management practice 135

4.6 Limitations and future research 136

CHAPTER 5 Mapping the signalling environment in sustainability-focused entrepreneurship and linking it to investment inputs: a topic modelling approach on company self-descriptions 137

5.1 Introduction 137

5.2 Theoretical framework 139
5.2.1 Sustainable technologies and information asymmetries 139
5.2.2 Signalling actions and sustainability-focused entrepreneurship 140
5.2.3 Task signals 142
5.2.4 Institutional signals 143
5.2.5 Task and Institutional signal consistency 144
5.2.6 Investment inputs 145
5.2.7 Research gaps 146

5.3 Data and methods 147
5.3.1 Dataset description 147
5.3.2 Keyword extraction using Textrank 149
5.3.3 Corpus pre-processing 151
5.3.4 Latent dirichlet allocation and Structural topic model 152
5.3.5 Estimation of the topic solution 155

5.4 Results 156
5.4.1 Text statistics 156
5.4.2 2-gram and 3-gram keywords extraction 158
5.4.3 Summary of the topic solution 161

5.4.4 Marginal effects 163
5.4.5 Mapping the dependence between topics....167
5.5 Discussion and implications 168
5.5.1 Implications for theory....169
5.5.2 Implication for practice....170
5.6 Limitations and future research recommendations....171
CHAPTER 6 Conclusions....173
6.1 Research findings 173
6.2 Theoretical implications 177
6.3 Practical implications 179
6.4 Limitations and future research 180
References....182
Appendix....214

List of tables

Table 1 A overall comparison across three studies 12
Table 2 Types and examples of unstructured data 18
Table 3 Overview of existing literature in rating prediction 45
Table 4 A comparison with other studies 48
Table 5 The comparisons of both advantages and disadvantage for 6 different algorithms 60
Table 6 The confusion matrix in the classification task 62
Table 7 The 15 topics and their top 7 loading words in the topic solution. 69
Table 8 Performance comparisons for all algorithms applied using document-level polarity in the classification task 70
Table 9 Performance comparisons for all algorithms applied using sentence-level polarity in the classification task 71
Table 10 Performance comparisons for all algorithms applied using topic membership in the classification task 72
Table 11 Three models' AUC Values for the classification task 76
Table 12 AUC Comparison among 15 topics' memberships in the classification task 78
Table 13 Two models construction with combination of sentiment and topic (3) membership 79
Table 14 Sample characteristics 102
Table 15 Contingency table of co-occurrence frequencies 106
Table 16 three types of metrics to measure the association. 107
Table 17 Top adjective-noun pairs 117
Table 18 Metrics to identify the collocation of "food" 120
Table 19 Metrics to identify the collocation of "time" 120
Table 20 Metrics to identify the collocation of "order" 121
Table 21 Metrics to identify the collocation of "delivery" 121
Table 22 Labels, distribution and top 7 loading words in the topic solution. 123
Table 23 Non-parametric Mann-Whitney U tests between the two types of restaurants 124
Table 24 Expected topic proportion across all brands from franchised restaurants 131

Table 25 Distribution of companies across 19 industries within the sustainable technologies group 148
Table 26 Summary statistics of the sample 149
Table 27 2-gram keywords 159
Table 28 3-gram keywords 160
Table 29 Labels, distribution and top 7 loading words in the 17-topics solution ... 162
Table 30 Research questions and findings 176

List of Figures

Figure 1 The inductive approach adopted in this dissertation.....10
Figure 2 Structure of this dissertation 13
Figure 3 The classification of unstructured data 19
Figure 4 The IMPACT cycle for unstructured data analytics 22
Figure 5 An example of customer reviews on JustEat 49
Figure 6 Distribution of rating scores within the reviews in our sample. The blue dashed line outlines the median rating score.....51
Figure 7 LDA model process using plate notation (Adopted by Blei et al., 2003)..55
Figure 8 Polarity score distribution for our sample 65
Figure 9 Selection of the number of topics (K) for identifying the topic solution. Optimal K is identified by the shaded area. 68
Figure 10 AUC comparison (left-hand side) between Model A and Model C as well as AUC comparison (right-hand side) between Model B and Model C 76
Figure 11 AUC comparison (left-hand side) between Model A and Model D as well as AUC comparison (right-hand side) between Model B and Model E 79
Figure 12 ROC curves for Model D and Model E after tuning the hyperparameters of XGBoost.....81
Figure 13 Relative difference of MAE and RMSE for 16 models compared with the baseline model.....82
Figure 14 SHAP summary plot 83
Figure 15 SHAP dependence plot for Topic #13 membership and its interaction visualisation with polarity score.....84
Figure 16 The rating score distribution for two types of restaurants 103
Figure 17 Research design for this study 103
Figure 18 An example of dependency structure 110
Figure 19 A graphical demonstration of STM process (adopted by Roberts et al., 2016) 111
Figure 20 Diagnostic values for topic solution with different K numbers.....114
Figure 21 Top nouns across the corpus.....115
Figure 22 Top adjectives across the corpus 116
Figure 23 A Network showing the collocation 122

Figure 24 The changes in the expected topic proportion from lower rating to higher rating for all restaurants 126
Figure 25 The changes in the expected topic proportion if the restaurant changed from being independent to being franchised 127
Figure 26 Principal component analysis: topic correlation plots for franchised versus independent restaurants 128
Figure 27 The interaction between rating score and restaurant type 129
Figure 28 Expected topic proportion distribution across brands 133
Figure 29 A graphical demonstration of STM (adopted by Roberts et al. 2016) 153
Figure 30 Diagnostic values for topic solution with 11-20 topics 155
Figure 31 Word histogram for adjectives 156
Figure 32 Word histogram for nouns 157
Figure 33 The changes in the expected topic proportions from smaller funding amount to larger funding amount 164
Figure 34 The changes in the expected topic proportions from less funding rounds to more funding rounds 165
Figure 35 The changes in the expected topic proportions from older company to newer company 166
Figure 36 Principal component analysis: topic correlation plots. Point size represents years of incorporation for major companies loading closer to the topic. 167

Acknowledgement

I would like to express my deepest gratitude to my primary supervisor, Nikolaos Korfiatis. Initially my external supervisor during my Master's dissertation at the University of Warwick, his diligence and intelligence have been a constant source of inspiration throughout my PhD journey, guiding me toward an academic career. He has been an exceptional supervisor and a supportive mentor. I am also grateful to my second supervisor, Dimitris Zissis, for his invaluable support in drafting this thesis.

Working with my coauthors, Antonios Karatzas and Dimitrios Dousios, has been a truly enjoyable and enriching experience. Through our collaboration, I have learned the importance of teamwork and organization in academic research. Plus, I would like to thank Konstantinos Chalvatzis and Konstantina Spanaki for their support in writing our papers. I extend my thanks to Aristidis K. Nikoloulopoulos for providing me the opportunity to teach undergraduate courses at the School of Computing Sciences.

I would also like to thank everyone at Norwich Business School for their support. The support from those around me during the COVID-19 pandemic has been invaluable. I am especially thankful to my best friend and colleague, Weijia Shao, for her encouragement and companionship throughout my PhD journey.

Completing this thesis would have been impossible without the unwavering support from every member in my family. I am deeply thankful for their love and encouragement over the years. In particular, I want to thank my father and mother for their continuous love and support.

Norwich, 2023

CHAPTER 1 Introduction

1.1 Research background

The environment of data availability and accessibility has significant changes as the result of the advancement of digital transformation in many organisations and nations. The widely use of digital technology thus producing associated exponential growth data could present a novel difficulty while enormous advantages within the field of management studies (Hannigan et al., 2019a). Data can be divided into structured data and unstructured data. Structured data could be well organised and easily queried while unstructured data is more complex and less predictable. Unstructured data including various forms of data such as text, images, and videos has been explored and utilised by both academics and industry (Balducci and Marinova, 2018). This is mostly due to its capacity to offer valuable insights that cannot be easily obtained solely from structured data.

The growing popularity of digital devices and access to the internet has led to a significant increase in the generation of data by humans (Dwivedi, 2020). The prevalence of this phenomenon can be observed in digital interactions that include information sharing, customer reviews, and various types of user-generated content (UGC) on platforms such as social media, social commerce, and electronic marketplaces such as Amazon and eBay (Donthu et al., 2021). In today's research, scholars are able to collect vast amounts of data from a variety of sources, including online forums, social media platforms, email communications, telecommunication devices, and sensor-based applications. The use of data is becoming increasingly prevalent in academic research, which is characterised by unstructured data and structured data.

Specifically, the service industry has acknowledged the significant importance of textual data by interpreting customer feedback, which is a type of unstructured data that can offer organisations valuable insights regarding their service quality, such as aspects that they can improve and customer attitudes towards their services. Within the service industry, particularly in industries such as hospitality, there has been a notable focus on comprehending the impact of customer feedback on company outcomes.

Among the service industry, food delivery industry has been emerging recent years, especially since the pandemic period. Food delivery has existed over a decade ago, since restaurant operators considered that food delivery could help to build strong relationship with customers and increase revenues. As technology eases people's daily lives, consumers are eager for faster delivery, accessible menus, personalised services and providing comments about their experiences. Restaurant operators desire for higher brand awareness, sales and customer loyalty, as well. Therefore, nowadays the online food delivery (OFD) sector is growing fast, and more consumers order food online through applications or websites because of its convenience and transparency. A lot of restaurants including franchised restaurants and independent restaurants put more effort on online takeaway delivery because of the temporary loss of eating-out market during the pandemic. The revenue of online food delivery (OFD) segment amounted to US$248.0 billion in the worldwide in 2020. With the impact of Covid-19 and national lockdown policy on eating-out market, the revenue of OFD market in the UK still shows a significantly increasing trend. UK owns the biggest OFD market in Europe, and the fourth biggest market over the world.

Customers have more tendency to read the contextual information within the reviews (Hu et al., 2008; Liu et al., 2019). The discovery of textual data embedded in customer reviews becomes necessary for businesses to exploit the informational value of customer feedback (Li et al., 2019). Review content plays an increasingly important role in the decision-making process before purchasing and consumers' purchase intention because of its persuasiveness and informativeness (Ruiz-Mafe et al., 2018; Liu et al., 2019; Schoenmueller et al., 2020). A customer review contains an overall rating (numerical score) indicating customers' overall satisfaction or dissatisfaction. The textual content associated with the rating score could explain the underlying rationale. Therefore, the association between rating score and review text could assist businesses understanding the reasons for different levels of customer satisfaction (Tan et al., 2016).

Apart from the customer-level unstructured data that offer customers' preferences and needs to companies thus improving the products and services, there is another form of unstructured data in several domains. The utilisation of firm-level data that is obtained from textual sources can play a crucial role in offering valuable perspectives on organisational strategies. The firm-level unstructured data also

contains valuable information that could be discovered from the textual content. For example, more and more companies provide their ESG report to disclosure organisation's operations and risks in environmental, social and governance (ESG) impacts. Text data has received considerable focus in the field of management research. This is primarily due to the fact that text data, encompassing a wide range of sources such as internal company reports. The textual data produced within the business context contains a significant amount of valuable information that can play a crucial role in contributing to the business value (Inmon and Nesavich, 2007). The organization could examine and analyse various types of textual data captured by them including online product descriptions, comments from suppliers or business partners. Frequently, unstructured data provides supplementary insights that might enhance the insights derived from structured data.

Textual data need to be pre-processed and cleaned before the further exploration. The information extracted from unstructured data could serve as a dimension, thus facilitating the integration of both structured and unstructured data inside the organisation (Baars and Kemper, 2008). Text mining is a computational method that has the ability to extract information and patterns from large textual data (Fuller, Biros and Delen, 2011), which is grounded in methodologies of data mining, machine learning, and natural language processing, which are employed to analyse textual data. The application of text mining has grown to the examination of trends and patterns in sustainability reporting (Zhou, Wang and Yuen, 2021). In addition, it has been utilised to determine the dominant components provided in corporate reports (Liew, Adhitya and Srinivasan, 2014).

1.2 Research questions

This dissertation attempts to discover the unstructured textual data in the business settings. To fulfil the research aim, there are three key questions that guide this research:

Research questions (RQ):

RQ1: How can be information from unstructured textual data extracted in business settings?

In the current corporate environment, the widespread existence of large volumes of data, especially unstructured textual data presents complex challenges however valuable information for businesses. How do we extract information from it? The first aim of this dissertation is to adequately extract the valuable information contained in textual data in a business setting using effective text mining techniques. The importance of this objective is significant because of the inherent challenges involved in the management and analysis of large datasets, which frequently contain extraneous and non-relevant material. This dissertation aims to adopt the adequate approaches from the related fields related to text mining such as natural language processing, information retrieval and machine learning in order to identify patterns and extract relevant information that is essential for business applications. It is a challenge to uncovering the information because of the distinctive characteristics of textual data such as the high dimensionality, sparsity and the existence of noise. The objective is to assist organisations in effectively managing unstructured data and utilising it to obtain a competitive advantage in the market.

RQ2: What is the role of unstructured data from the business perspective?

The second objective of this dissertation aims to provide a comprehensive understanding of the vital importance and multitude of benefits of unstructured textual data in the context of business. Unstructured textual data possesses certain characteristics such as high dimensionality, sparsity, and noise. This type of data serves as a great source of latent information, containing crucial insights that are essential for the success of organisations (Müller et al., 2016). The utilisation of this particular type of data, which is mostly sourced from company disclosures, project documents, and customer interactions, has the potential to reveal complex patterns and significant insights. Consequently, it may greatly support the decision-making process by providing valuable information. The application of advanced analytical methods, such as Natural Language Processing (NLP) and Machine Learning (ML), is crucial in converting unorganised data into organised, practical knowledge. This enables businesses to derive comprehensive insights and improve their strategic perspective (Adnan and Akbar, 2019). The amount of unstructured data that may be utilised depends on the effectiveness of the analytical methods used and the contextual

significance of the collected information. The second objective of this dissertation attempts to examine how business operations could be enhanced with the help of unstructured textual data.

RQ3: How can unstructured data be linked with the structured data and the business context?

The third objective aim of this dissertation is to investigate how the unstructured data is linked with the structured data from business and how it can be quantified. The incorporation of both structured and unstructured data is crucial for organisations to efficiently acquire extensive and highly dependable datasets consisting of structured attributes and outcomes. These datasets can then be utilised as a fundamental component in decision making procedures. Besides, textual data is highly related with the context, by discover the relationship of information within textual data and the business context, the association could be discovered quantitatively. This objective is to link the information extracted from unstructured data with the structured data from business context and quantity the connections.

1.3 Research methodology

In this section, the research philosophy, research approach and techniques as well as the dataset will be discussed in detail. The researchers in social science are actively involved in discussions about how various text mining tools can be used (Howell, 2012). They have critically analysed the epistemological assumptions, ontological assumptions and metatheoretical assumptions, which corresponds to the assumptions about the nature of knowledge, the nature of reality and the capacities and limitations of scientific theories respectively (Ignatow and Rada Mihalcea, 2017). The research philosophy can be considered as the fundamental nature of how the research is shaped and developed. There are assumptions under each study, with each showing the underlying assumptions and how researchers obtaining the knowledge (Žukauskas, Vveinhardt and Andriukaitienė, 2018).

Positivism is a perspective of research philosophy which is based on the belief in objectivity, which mean the social world could be interpreted in an objective way. This perspective is based on the ontological position that posits the existence of a

solitary, readily accessible reality. Positivists tend to use the way of exploring natural world to understand the social world by identifying the cause-effect relationship observed in phenomena, thus used for predicting in the future in the social world since the reality is not affected by the context. Positivism prioritises the objective examination of 'facts' that are perceived to exist independently from individual emotions or social ties to the world. The researcher should act as an objective analyst who can work independently doesn't include his/her own personal values (May and Perry, 2022). According to Hutchineson (2004) *"Positivists view the world as being 'out there', and available for study in a more or less static form"*

Positivist research is commonly associated with the generation of quantitative data. The positivist epistemological viewpoint is closely aligned with the quantitative representation and study of social phenomena. This perspective operates under the assumption that social reality maintains constant features across various settings and time. Consequently, it is possible to identify and categorise specific attributes as separate variables, which could own different values, which can be quantified using numerical scales. Positivist researchers use quantitative data in their research and theoretical formulation. These data are obtained through rigorous actual experiments, as well as less rigorous quasi-experiments. The closed ended questionnaires are also conducted as well as the standardised tests. The methodology is grounded in positivist research paradigm and uses statistical tools to transform observations into quantifiable data that can then be empirically evaluated. Building on the foundations of positivist epistemology, this perspective views data as objectively observable and free from any form of bias.

Positivist research mostly relies on the utilisation of experimental methodologies. This entails formulating hypotheses that propose causal relationships between certain occurrences. Following the collection of empirical evidence, the acquired data is subjected to meticulous analysis, resulting in the development of a theoretical framework that demonstrate that how the independent variable will impact the dependent variable. Experimental techniques under the framework of positivism frequently entail the manipulation of environmental variables in order to ascertain the causal causes underlying observed changes. The assumption is made that the fundamental principles governing scientific inquiry remain constant, irrespective of fluctuations in local circumstances. Consequently, it is believed popularly that the

results obtained from controlled environments may be extrapolated to real-world situations, regardless of the specific context (May and Perry, 2022).

The origins of quantitative textual analysis can be traced back to earlier efforts to study newspaper content and show a strong correlation with the basic concepts of content analysis. The process involves the quantitative analysis of textual data in order to identify dominant themes and messages conveyed. Roberts (2000) highlights the unique characteristics of the study, emphasising its well-defined textual scope and the ability to quantitatively determine degree of error in the analysis. Within this particular paradigm, scholars aim to measure the occurrence of particular themes or keywords, analyse their connections through semantic structures, and examine their position within a vast network of thematic connections.

Quantitative methodology is firmly grounded in the positivist worldview and is widely accepted as a scientific technique (Jr and Unrau, 2010). Similarly, this particular approach seeks to achieve objectivity through the measurement of behaviours and opinions, allowing researchers to provide facts in a descriptive way rather than engaging in interpretation (Rahi, 2017). The study relies on the use of conceptual framework influenced by the positivist perspective, which emphasises the establishment of causality, verification and generalisability.

In recent years, there has been an increase in the usage of terms to describe the numerous alternatives for analysing message content with computer algorithms. Computers have been instructed to obtain and process messages for diverse objectives, thus greatly improving the potential of computational analysis applied in content analysis. The utilisation of computer-aided text analysis (CATA) has facilitated the integration of qualitative and quantitative analysis techniques in analysing the text, enabling both analysts and non-analysts to employ these approaches and processes. CATA refers to a method that utilises computer software to methodically and objectively identify predetermined attributes inside text, with the purpose of drawing inferences from the text. Content analysis is a methodical approach to quantitatively assessing textual or symbolic elements (Krippendorff, 2018).

The quantitative content analysis is getting to attract more attention. In the quantitative content analysis, there are two primary approaches available within the coding process: human coder and computer-assisted coding, also referred as CATA.

This approach offers an expanding range of possibilities for doing automated analyses for text using computer. The advent of technology-driven analysis has facilitated the popularity of the application of text analysis in various sectors and domains of research.

The terms "content analysis" and "text analysis" are frequently employed synonymously. The phrase "text mining" has been widely used to encompass all the underlying mechanisms involved in text analysis (Sebastiani, 2002). Text mining and content analysis rely on text as the primary medium for extracting content and gathering information. Recent work has provided a comprehensive account of the similarities and complimentary aspects inherent in the two approaches to. Yu, Jannasch-Pennell and DiGangi (2011) claim that text mining is congruent with content analysis while Lin, Hsieh and Chuang (2009) believed that using text mining could be used as an enabler of content analysis.

In the field of text analytics, it is essential for every instance of acquired textual data to possess a distinct identifier which enables convenient referencing to individual records. These occurrences are commonly known as documents. In certain instances, the records may be not physical documents, such as individual tweets or consumer reviews. All instances comprise a collection of documents.

Similar with content analysis, the scope of analysis in this dissertation is a major decision which means the unit and granularity of analysis need to be decided. The unit of analysis encompass a range of textual elements, including but not limited to documents, sentences, phrases, words, or characters. Text data analysis at the word or character level is more prevalent in the field of natural language processing as compared to text analytics. In the field of text analytics, the primary emphasis is generally placed on the level of individual documents. In this dissertation, text is analysed hieratically from a variety of levels

In addition, it is vital to ascertain the specific information inside the documents that holds the highest significance. This may encompass document categories, semantic implications, or the content. When the objective is to classify or categorise materials, the approach employed will differ from that used to ascertain the semantic information of these texts. Nevertheless, the dissertation is interested in the content of textual data including the semantic meaning and sentiment within in the

text. semantic information from documents provides insights on the interrelationships between words within the documents, including co-occurrences (Griffiths, Steyvers and Tenenbaum, 2007). Sentiment information pertains to the subjective aspects of text, encompassing emotions, feelings, and the polarity derived from the words contained within the documents.

Unlike quantitative or numeric data, unstructured data is not measurable which requires other techniques. Content analysis could be a suitable methodology for unstructured data as it is "*a research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use*" (Krippendorff 2012, p.18). The approach we employ for text analytics is guided by the rigorous principles of content analysis theory. Our approach to text mining and analytics is based on the content analysis framework outlined by Anandarajan, Hill and Nolan (2019).

Content analysis allows using either deductive or inductive approach to make inferences and draw conclusions. The deductive approach refers to a logical reasoning process, which is characterised by its reliance on pre-existing theories, as opposed to being derived from empirical facts. The present approach relies on a pre-existing hypothesis that has been previously established and proven by prior study. The research method generally entails the collection of empirical data through the systematic observation of phenomena, which is guided by a pre-established conceptual and theoretical framework. In the filed of scientific inquiry, scholars typically endeavour to empirically evaluate a theoretical framework by collecting fresh data from individuals involved in the study. The collected data would be analysed statistically in order to draw meaningful conclusions. The process of data analysis often follows a deductive approach, commencing with the formulation of a hypothesis, which is subsequently tested and either confirmed or disproven through the examination of statistical evidence. The primary objective is to quantitatively measure, regulate, predict, develop legal frameworks, and ascertain causality (Cohen, Manion and Morrison, 2017).

The inductive approach is a methodological approach that involves drawing general conclusions based on specific observations, which is characterised by the researcher deriving theory from observations obtained during the research process. The act of drawing conclusions or forming inferences about particular subjects or

variables based on observable data might be seen as a key aspect of this phenomenon. The inductive approach uses an open coding system where the analysis is influenced by the data rather than a pre-existing theory. Therefore, in general this dissertation follows an inductive approach as shown in Figure 1, which is adapted from the content analysis framework and provides an overview of the induction approach that is adopted in this thesis. The text mining techniques involves several various techniques related to the relevant fields including Natural Language Processing, machine learning, etc.

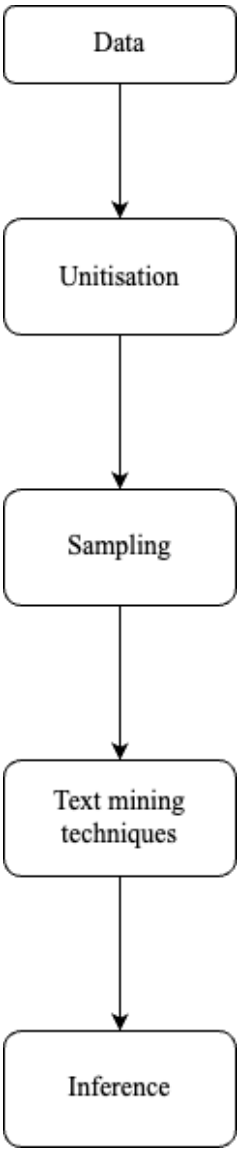

Figure 1 The inductive approach adopted in this dissertation.

In order to address the three research questions, we conduct three separate studies to answer these questions. The following table demonstrates the datasets and methods used in three studies.

Table 1 A overall comparison across three studies

	Study I	Study II	Study III
Dataset	Unstructured data and structured data	Unstructured data and structured data	Unstructured data and structured data
Dataset description	Online customer reviews (ratings and review content)	Online customer reviews (ratings and review content) Restaurant-relevant information	Companies' self-descriptions Companies funding information
Level of data	Customer-level data	Customer-level Firm-level data	Firm-level data Industry-level data
Granularity of analysis	Sentence level Document level	Word level 2-gram Document level	Word level 2-gram 3-gram Document level
Techniques	Topic models Machine learning techniques	Topic models NLP techniques Regression	Topic models NLP techniques Regression

In three studies, datasets contain unstructured data (textual data) and structured data (numerical data). In study I, online reviews including review ratings and review content are collected from JustEat, which is the main online food delivery platform in the UK. In study II, except from the online reviews, the restaurant-relevant information from JustEat is also collected. In study III, the dataset is collected from Crunchbase, which covers the companies from sustainability industry with its self-descriptions (unstructured data) and funding status (structured data). The granularity of analysis covers various levels: word level, 2-gram, 3-gram, sentence level and document level. A variety of text mining techniques has been applied in three studies which covers the field of Natural Language Processing, machine learning, etc.

1.4 Research outline

The thesis is organised as follows. Chapter 1 starts with an introductory overview of the context and rationale of undertaking this study. Research questions are followed after that. Then the research philosophy, approaches, framework are discussed in detail.

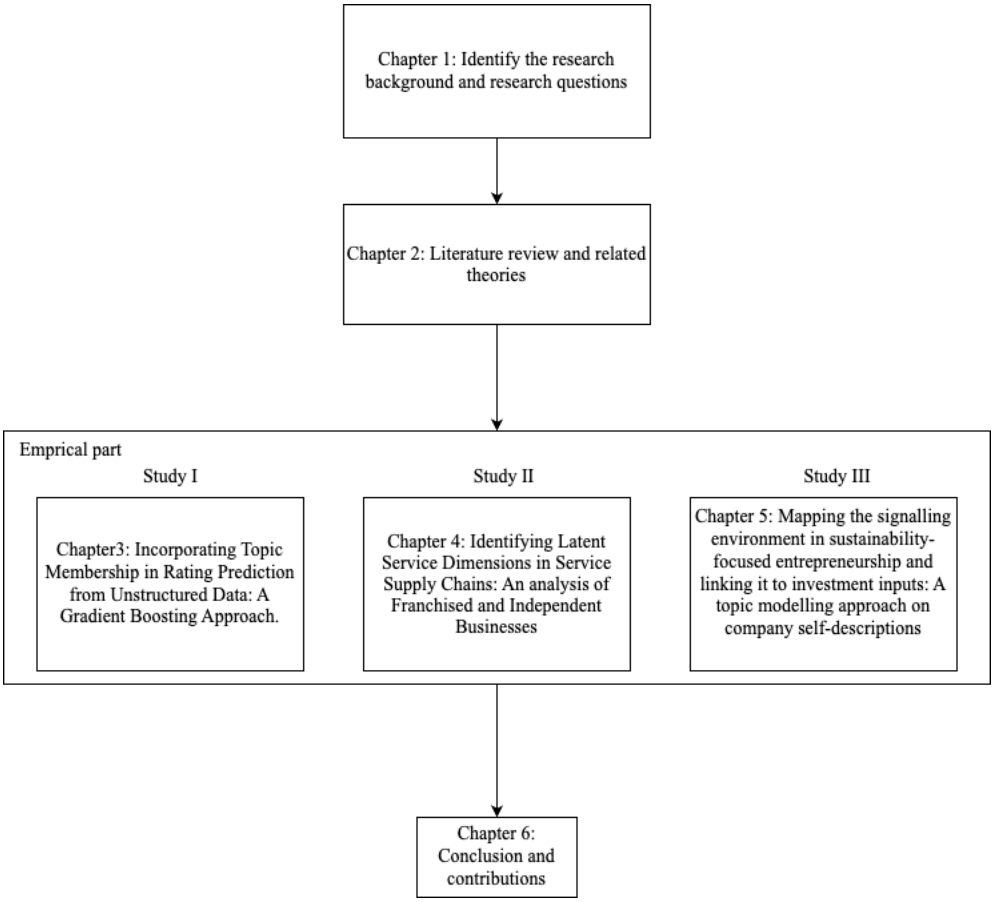

Figure 2 Structure of this dissertation

Chapter 2 presents a comprehensive review of unstructured data and unstructured data analytics. More specially, we focus on the unstructured textual data. The characteristics of text and the granularity of text are demonstrated. Then the adaptation from content analysis to text analytics is also illustrated.

Chapter 3, 4, and 5 are three relatively independent empirical studies with each representing an introduction, literature review, data and methods and findings. Chapter 3 presents Study I entitle "Incorporating Topic Membership in Rating Prediction from Unstructured Data: A Gradient Boosting Approach." Study I demonstrates the rating prediction using customer feedback. It presents a dynamic model that incorporates values from topic membership, an outcome variable from

Latent Dirichlet Allocation, with sentiment analysis in an Extreme Gradient Boosting (XGBoost) model used for rating prediction.

Chapter 4 presents Study II entitled “Identifying Latent Service Dimensions in Service Supply Chains: An analysis of Franchised and Independent Businesses”. Study II concentrates on the discovery of online reviews from customers. It covers a wider range of analytical methods from three levels: word-level, phrase level and document level. the various methods are utilised to discover the different level of information from online review thus developing knowledge for business.

Chapter 5 present Study III entitled “Mapping the signalling environment in sustainability-focused entrepreneurship and linking it to investment inputs: A topic modelling approach on company self-descriptions”. It suggests a topic modelling approach to extract latent signals from the self-descriptions of companies in the sustainability sector. It also illustrates how the latent signals are connected with the organisational context and the funding outcomes.

Chapter 6 starts by summarising the findings from three empirical studies in chapter 3,4 and 5. This is followed by a discussion on the theoretical implication and implications for practices, the limitations of this thesis and the directions for future research.

CHAPTER 2 Literature review

2.1 Unstructured data

2.1.1 Unstructured data and structured data

The categorisation of data in corporate environments is a multifaceted task amenable to a variety of segmentation strategies (Inmon and Linstedt, 2015). An example is the division of data into structured and unstructured data. The distinction between structured and unstructured data in data management and analysis highlights fundamental differences in their inherent organisation and accessibility.

Structured data adheres to predetermined formats and durations, which facilitate storage and analysis with a strong sense of organisation. This organisational feature underscores the data's arrangement in distinct structures, rendering it more responsive to inquiries designed to retrieve information for organisational purposes (Eberendu, 2016). Structured data consists of information characterised by a consistent and predictable data format. This category of data is typically managed by a database management system (DBMS), which includes the components records, attributes, keys, and indexes. Structured data is distinguished by its well-defined and predictable attributes, which are administered by a sophisticated infrastructure (Chen et al., 2009). Relational databases, exemplified by platforms like Access or structured query language (SQL), provide a paradigmatic illustration of structured data (Duan and Xiong, 2015). The data within these databases is meticulously organised and consists of numerical values, dates, and composite entities of alphanumeric sequences or text. The natural cohesion of the database architecture allows for easy navigation through uncomplicated and direct search algorithms, often based on data type and content attribute criteria. Consequently, the structured environment expedites the retrieval of the majority of data units. Structured data illustrate information characterised by a high level of organisation, allowing for seamless incorporation within relational databases and facilitating exploration via conventional search engine algorithms and search operations.

Unstructured data is defined by Isson Paul Jean (2018) as *"The term refers to data that does not have a predefined data model and/or does not fit well into traditional relational database tables. Unstructured data typically has no identifiable*

structure and may include bitmap images/objects, text, audio, video, and other data types."

Unstructured data presents itself as unpredictable data in which there is no structure discernible by computational systems. Accessing unstructured data is typically a laborious process, requiring sequential queries (parsing) through extensive data processes to locate particular data units. In contrast, unstructured data represents a different paradigm, characterised by a lack of predefined schema and a tendency towards a lack of systematic order. Consequently, the process of curating unstructured data is frequently resource- and time-intensive, due to the inherent complexity of its assimilation.

The famous example of electronic communication (email) exemplifies unstructured data. Despite the fact that a corporate human resources manager's inbox may appear structured based on temporal or volumetric characteristics, a truly structured representation would require meticulous categorization by precise subject and content. However, such an approach is impracticable, as the lexicon of human interaction frequently incorporates multiple themes, even within emails that are ostensibly focused (Zohuri, Mossavar-Rahmani and Behgounia, 2022). Nevertheless, the structured nature of spreadsheets corresponds with the characteristics of structured data, allowing for rapid perusal within relational database systems due to their alignment. The underlying problem posed by unstructured data is its volume; a substantial portion of business interactions take this form, necessitating substantial investments of resources for discernment and extraction, similar to the algorithms underlying web search engines. Existing data mining techniques frequently neglect a substantial portion of latent information that, if subjected to judicious analysis, could yield potentially transformative insights of strategic importance (Inmon and Linstedt, 2015)

However, There is an intermediate data format that occupies a position between the dichotomous domains of structured and unstructured data. The intermediate form has a noticeable structure, but it deviates from the usual assumptions typically associated with structured data. As a result, this particular structure does not lend itself well to direct conversion into the tabular format characteristic of relational databases. This type of data is known as semi-unstructured

data. In some way, semi-structured data could be considered as a data type in between structured data and unstructured data (Schroeck, 2012).

Semi-structured data bears similarities to structured data but with structural differences (Roberts, 2000). In this case, data retains its inherent structure, but format consistency varies between data items. In the context of semi-structured data, the establishment of a structural framework is usually achieved through the use of tags or similar markers. These tags have the dual function of delineating and separating individual fragments of data. These markers, in turn, allow the creation of records, often characterised by hierarchical relationships, similar to a tree structure. It is important to note that the schema that governs semi-structured data is not always predetermined, which allows for a significant degree of flexibility. As a result, it is not necessary for different occurrences of the same entity to have identical characteristics. The ability to adapt also allows for the inclusion of redundancy and non-atomic values (Ling and Dobbie, 2004). XML and JSON are commonly seen as prototypical examples of semi-structured formats.

2.1.2 Types of unstructured data

Text emerges as the most pervasive manifestation of unstructured data, which encompasses a multitude of manifestations and variations. However, it is essential to emphasise that unstructured data encompasses diverse forms and modalities in addition to textual representation. The effort to construct a mechanism for unstructured data analysis is relevant across multiple levels of business operations for enterprises operating in diverse industries. The discovery of such a mechanism has the potential to alleviate the financial burden imposed by the complex management of unstructured data within an organisational framework.

The unstructured data including video and image has been utilised by several researchers in the business studies. One study conducted by Lu, Xiao and Ding (2016) presents an automated and scalable system which aims to offer personalized recommendations to shoppers by utilizing real-time in-store videos. The main goal of the system is to improve the shopping experience for customers and therefore enhance product sales. Pantano, Dennis and Alamanos (2021) developed a novel system using machine learning algorithms to identify consumers' emotional responses from images.

Instead of examining emotions from an internal perspective, they focused on the detect of consumer emotion using unstructured data (i.e., image).

Table 2 Types and examples of unstructured data

Unstructured Data Type	**Examples**
Text	E-mails Chats Tweets Facebook posts and updates Other social media Customer feedback Product reviews Customer narratives Voice of customer (speech to text) Open-ended questions (satisfaction survey/survey) Job posting description
Image	Pictures Bitmap
Audio	MP3 Phone call Voice recognition
Video	Eye tracking system Produce description video

The possibility of applying manual or automated techniques that facilitate the assignment of identifiers to formerly unstructured data reveals a route to creating structure. Unstructured data can be transformed into a semi-structured layout that includes interpretive indicators comparable to those discovered in structured and semi-structured frameworks. The idealistic prospect of instantly transforming

unstructured data into a structured format, thereby speeding up the extraction of actionable insights, requires nuanced consideration. Structured data exemplify a computational language similar to machine code, allowing computers to interpret and manipulate data more efficiently (Losee, 2006). In contrast, unstructured data are, to some extent, designed for human comprehension, taking into account the nuanced and multifaceted nature of human communication, which may not adhere to strict database formatting.

The structural characteristics of individual data elements are along a continuum from highly unstructured to highly structured. The location of a data unit along this continuum indicates the ease with which structure can be incorporated into the data to make it amenable to quantitative analysis intended to gain generalizable insights (Balducci and Marinova, 2018). There are some data units, which contain a number of concurrent data points. For instance, video data contains non-verbal signals, spoken words and acoustic vocal cues, which requires the assignment of values to them by the researchers manually or automatically before quantitative analysis is employed. On the other hand, the structured data are data units that require relatively little effort or no effort to prepare for analysis.

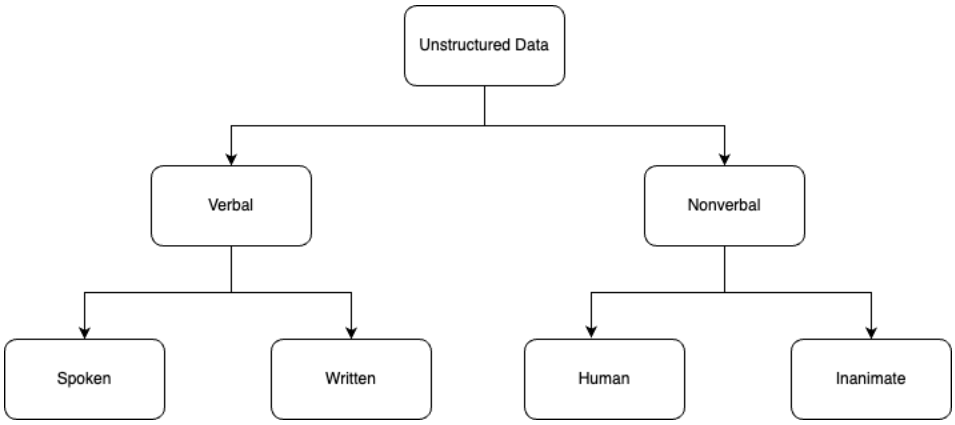

Figure 3 The classification of unstructured data

2.1.3 Characteristic of unstructured data

Non-numeric is the first characteristic of data units classified as highly unstructured data. These data units do not have predetermined numeric identifiers for

the investigated constructs, thus requiring manual or automated coding by researchers prior to analysis. For example, if researchers want to determine the level of customer affect portrayed through non-verbal cues during a service interaction, they need to identify the nonverbal signals that convey positive, negative, and neutral affect with varying degrees of intensity. The frequency of occurrences and the level of affect expressed through each cue require a lot of effort to transform them into more structural data. Highly structured data has a pre-established numeric representation for the targeted construct. Even if numerical values are not explicitly present, researchers can assign them easily, often by utilising categorical data classifications (Balducci and Marinova, 2018).

Moreover, another notable feature of highly unstructured data is their multifaceted nature. Each unit expresses a diversity of attributes, each of which provides different and unique information (Korfiatis et al., 2019). This attribute allows researchers to focus on and examine specific aspects according to their research objectives. For example, voice data naturally encompasses a range of attributes including pitch, speech rate, and volume. Each of these aspects conveys unique information about the speaker, such as emotional subtleties and persuasiveness. In contrast, the facet structure of highly structured data is typically monolithic, which represents a solitary unique piece of information. Consider the net promoter score, which can be interpreted from numerous perspectives, such as service quality or customer satisfaction. However, this data unit is dominated by a single feature: the numerical score. Whether to interpret this data unit as indicative of service quality or customer satisfaction depends on the researcher's perspective, rather than a fundamental difference in selected aspects inherent to this data unit.

What’s more, highly unstructured data elements exhibit concurrent representation, with multiple facets providing distinct information within a single data element. This results in the capability of unstructured data to embody multiple phenomena simultaneously. This innate feature allows researchers to explore numerous research objectives with the help of the exploration of the concurrent flow of the aspects in unstructured data. For instance, the textual data from the commutation between a customer and a sales representative has several unique facets such as syntax and semantics that occur at the same time, which could provide diverse information. In contrast, highly structured data is uni-faceted, causing a

fundamentally non-concurrent mode of representation. Despite the fluidity existing in highly structured data (such as high frequency time series data), the single data instance only reports one facet at once (Balducci and Marinova, 2018).

2.1.4 Unstructured data analytics

The human cognitive ability demonstrates the capability to understand written material, whether it is presented in the form of a sentence, a paragraph, or an entire document. Furthermore, the field of computational processing can be used to provide approaches with similar capabilities. Achieving a thorough understanding of information, whether presented through text or images, necessitates the conversion of unstructured data into quantitative representations (primarily numerical values). The computing technology must be capable of deciphering, interpreting and grasping the hierarchical structure of linguistic components including phrases, sentences and paragraphs. Through the use of mathematical algorithms, the meaningful information could be extracted from the unstructured data while eliminating unnecessary noise and redundancies. Despite significant progress over the past decade, it remains a challenging task to provide a computer with a level of textual understanding equivalent to that of a human Isson Paul Jean (2018) proposed a framework to create business value from unstructured data. The IMPACT cycle offers a structured framework for deriving actionable insights from both structured and unstructured data sources. The subsequent section outlines the IMPACT cycle's sequential stages:

Identify the question. In a non-intrusive manner, the framework promotes collaboration between the analyst and business partner to pinpoint key business problems that necessitate resolution. This collaborative exploration establishes the groundwork for understanding the temporal requirement and extent of effort required to provide responses.

Master the data. It requires a thorough gathering, examination, and integration of relevant information to answer the identified business problems objectively. The condensed data is subsequently transformed into clear and easy-to-understand visual representation like diagrams, charts, tables, and interactive platforms in order to guarantee accessibility and clarity.

Provide the meaning. From the analysed data, this stage involves articulating clear and concise interpretations that are relevant to the important business questions

previously indicated. The key to this phase is connecting the data insights with the overall business context.

Actionable Recommendations. This section involves crafting business recommendations via the interpretation of data. Even if these proposals contrast with established norms, their responsiveness is generally less complicated than developing entirely new concepts. Whenever feasible, these recommendations come paired with estimated monetary values that underscore the potential for increased revenue or reduced costs.

Communicate Insights. This phase demonstrates a diverse communication strategy aimed at disseminating insights throughout the organisation. The strategies encompass interactive tools, recorded presentations, knowledge-sharing sessions and executive memorandums, with the objective of maximising the distribution of insight.

Track Outcomes: The final phase establishes a process for evaluating the impact of the obtained insights. It is recommended that the influence of insights be continuously monitored and that follow-ups with business counterparts are conducted to clarify the outcomes of implemented actions. This iterative process involves retrospective evaluations of actions taken, their consequences, and the emergence of new central questions that require analytic intervention (Isson and Harriott, 2012).

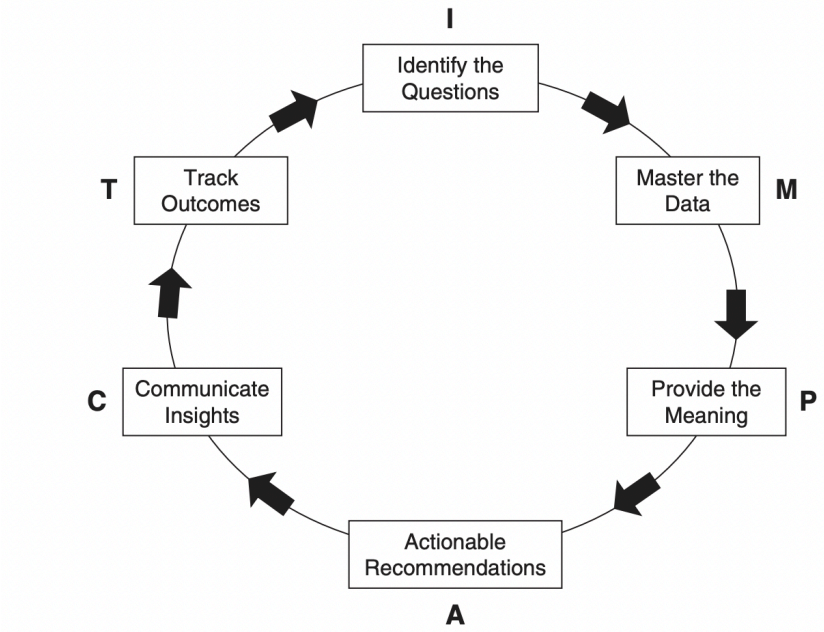

Figure 4 The IMPACT cycle for unstructured data analytics

2.2 Textual data

2.2.1 Definitive characteristics of text

Text possesses no reader-independent attributes by definition. Identifying an object as text extends an invitation or even an implicit obligation to read it. Designating certain indicators as data requires the unwavering recognition of these indicators as an indisputable foundation for subsequent conceptualisations. Therefore, texts, messages and data come into existence based on individual conceptual engagement with them.

The presence of a reader is necessary for the existence of a text, while an interpreter is needed for the existence of a message. Likewise, data only take shape through the involvement of an observer. In the context of content analysis, researchers with methodological expertise and intimate familiarity of the texts being examined are responsible for formulating analysis, guiding researchers in explicating textual components, and interpreting the results. This interpretation process always considers the comprehension of others. Importantly, it should be noted that a text inherently lacks characteristics for which its meanings are derived through individuals' interpretations. It is essential to acknowledge that the interpretation of texts by ordinary readers and content analysts differ.

Texts lack inherent single meanings that can be conclusively determined or exhaustively described in isolation or in direct correlation to their sources. Similarly, texts can have multiple interpretations, indicating that signs embedded within texts can comprise a variety of distinct designations, and data can be subject to a variety of analytical approaches. Considerations may vary from numerical assessments, such as character, word, or sentence amounts, to categorisations of expressions, analyses of metaphorical elements, explanations of logical compositional structures, and evaluations of associations, meanings, references and commands within the text. Additionally, interpretations may span various fields, including psychiatric, sociological, political, or poetic dimensions, each providing a valid but distinct viewpoint (Krippendorff, 2018). This multitude of analytical paths can be daunting for analysts lacking the necessary training.

Although it would be helpful if there was agreement on an author's intended message or the communal interpretation of a text, it is rare. Requiring analysts to

focus only on "common" or "shared ground" would significantly restrict the empirical scope analysis. It could limit this analytical methodology to scrutinising solely the fundamental or evident facets of communication. Alternatively, it could restrict the analysis to a tight-knit group of message creators, recipients, and examiners who share a particular perspective. When using content analysis to process the text, the bedrock of content analysis is founded on the notion that analysts can construe texts distinctively from other readers. This variability is not a flaw, but an inherent feature that enables content analysis to achieve its goal. The effectiveness of content analysis can be hindered if specialist interpretations neglect the diversity of textual applications by specific reader populations or individuals. This issue is exacerbated when content analysts do not clarify the criteria utilised to confirm their findings.

In contrast to the conventional notion that messages "contain" or inherently "possess" meanings or contents, the crux of meanings or contents goes beyond the limitations of the texts themselves. Communication's primary trait is its potential to communicate information, evoke emotional responses, and prompt behavioural alterations. Texts have the capacity to provide readers with information about remote events, descriptions of extinct entities, insights into the thoughts of other individuals, and available courses of action. Like symbols representing absent objects and narratives guiding their audience through imagined worlds, texts can connect current readings to elements beyond the immediate context.

Researchers must possess the ability to conceptualise and articulate the interrelationships between external elements, which may include solely mental constructs, past or future events, or hidden causal factors. This highlights a fundamental limitation that is endemic to computer text analysis since it is necessary to look beyond the materiality of texts. In consequence, they must investigate how non-analysts utilise these texts, what insights are provided, and what ideas and actions are inspired. While computers can be programmed extensively to manipulate character sequences, their operations are confined by the conceptual boundaries established by their programmers. In the absence of human intelligence and the inherent ability to grasp and deduce meaning from texts, computer text analysis is restricted to data processing. Computers lack their own environment and function within the contextual confines of their users' worlds without true comprehension.

The meaning of texts is based on specific contexts, discourses or intentions (Trilling and Jonkman, 2018). Texts can be interpreted in multiple ways, but this doesn't make the task of content analysts pointless. Messages are always part of certain situational contexts, texts are understood with certain objectives in mind, and data only becomes relevant in relation to specific issues. Variations in interpretation can result from analysts' choice of contextual lenses for interpreting discourse. However, multiple interpretations do not necessarily preclude the possibility of reaching consensus in a given context. It requires a framework that can convert unstructured perceptual information into organised materials and forms the conceptual foundation for producing valid conclusions. It is crucial to note that interpretive frameworks used by unaided observers, viewers, and readers can differ from the contexts at play when interacting with sensory data, textual characters, and incoming messages. As a result, distinct reader groups with different analysts can produce widely varying results when scrutinising the same corpus of texts. To guarantee reproducibility of analysis, the context needs to be clarified.

2.2.2 Features of text at different levels

The choice of configuration for the attributes that represent the characteristics of a document depends on the particular problem to be solved. For example, the use of single characters as attributes may be advantageous in the context of spam filtering. In this context, the presence of certain special characters or excessive repetition of identical characters can act as indicators of spam. On the other hand, in the context of an information extraction task aimed at obtaining meaningful semantic information for human users, the concept of character-level properties becomes irrelevant. In such situations, the use of words or abstract concepts at a higher level of complexity plays a crucial and vital role (Zizka, Darena and Svoboda, 2019). In most cases, it is possible to establish a clear classification of feature types based on the characteristics of the data as shown below:

Characters: The process of choosing features to represent textual data is closely connected to the particular problem being addressed. In specific situations, the utilisation of individual characters, character bi-grams, tri-grams, or more extended n-grams may be perceived as the most inclusive method of representation.

Nevertheless, the utilisation of characters as features is not commonly employed, partly because of their restricted ability to communicate semantic information. The process of extracting meaning from individual characters can be somewhat complex, particularly when the contextual information, such as the character's position within the text, is ignored. This approach is usually known as a "bag-of-characters" representation.

Words: words are identified as the most inherent and instinctive characteristics that may be derived from textual information. A word, which can be described as a series of symbols, commonly consisting of letters, is utilised within sentences in a certain language. A term exhibits intrinsic semantic, grammatical, and phonological consistency and is typically separated from neighbouring terms by spaces in written format (Dixon and Aikhenvald, 2003). The typical word structure adheres to the linguistic conventions commonly seen in Western languages. However, in many languages such as Chinese or Arabic, word boundaries are frequently not indicated by spaces, making it more difficult to visually identify these borders.

The use of dictionaries is crucial in determining the existence of words in a particular text. In general, dictionaries designed for natural languages consist of a comprehensive collection of words, commonly including the lemma, which represents the base form of each word. As an illustration, the lemma for several forms of the verb "do," such as "does," "doing," "done," or "did," is unequivocally "do." Hence, the nonexistence of a specific character sequence inside a lexicon does not definitively classify it as an illegitimate term. In addition to linguistic units, certain textual components, such as numerical values, temporal references, monetary symbols, and punctuation marks (e.g., commas, periods, question marks, quote marks, etc.), might be regarded as potential attributes, depending on the particular analytical objective.

The significance of these factors may exhibit variability contingent upon the contextual framework. For example, when a numerical value is assigned to denote the cost of a product, its magnitude plays a crucial role in classifying the object as either high-priced or reasonably priced. On the other hand, in specific situations, it may be adequate to have knowledge that a text has a particular expression connected to pricing. Nevertheless, it is crucial to recognise a constraint linked to features based

on words, specifically, their reliance on context. The potential interpretation of a word can be obscured in the lack of contextual information.

Terms: In the field of textual analysis, terms play a crucial role as essential components, encompassing both individual words and multiword formulations formed from the text. Upon analysing the previously described fabricated news headlines outlined in the preceding section, it becomes evident that certain terms such as "Las Vegas," "ice hockey," "FIFA World Cup," and "the National Lottery" may be readily discerned. Terms serve the purpose of reducing the inherent ambiguity that arises from individual words by combining many words into a unified entity. The process of term extraction is closely related to the Named Entity Recognition (NER) problem, which is a specific area within the broader field of Information Extraction tasks. Named Entity Recognition (NER) is a computational task that involves identifying and classifying specific types of information entities inside a given text. These entities can include personal names, organisations, geographical places, as well as quantitative expressions denoting time references, dates, monetary values, and percentage expressions (Nadeau and Sekine, 2007).

Concepts: Concepts are identifiable characteristics that are created by analysing the textual content. These characteristics are determined by the presence of words, sentences, or more complex linguistic structures. The clear manifestation of these concepts is not necessary inside the text. In the realm of news headlines, it is possible to deduce notions such as "sport" or "gambling," for example. Concepts, when used in conjunction with terms, possess a significant amount of semantic depth, which aids in the more efficient management of linguistic complexities such as synonymy, polysemy, hyponyms, and hypernyms. Nevertheless, the process of extracting these entities usually requires a significant amount of computer resources, and their significance is often dependent on the particular field of study. The process of accurately assigning concepts often involves comparing and verifying information from external sources. These sources can include the knowledge and expertise of specialists in the field, established ontologies or lexicons, or a carefully curated collection of annotated training documents used to train algorithms specifically designed for concept assignment.

In the field of natural language processing, it is common to group characters or words into so-called n-grams. Specifically, a character n-gram refers to a

consecutive sequence of n characters. For example, in the given text "A few character 3-grams created from this sentence", the character 3-grams include "A f", "fe", "few", "ew", "w 3", and other similar combinations. Similarly, word 2-grams within a given sentence include combinations such as "word 2-grams," "2-grams of," "of this," and similar phrases. Character n-grams are able to represent the whole words and also some word categories (including ‘ed’, ‘ing’), which are advantageous distinctly than the word n-grams in terms of diminishing the concern about dimensionality because the number of possible character combinations is comparatively smaller than that of word combinations (Kanaris et al., 2007). Therefore, character n-grams has been applied in several domains and examined to be effective in language identification (Brown, 2013) and authorship identification (Houvardas and Stamatatos, 2006). In contrast, word n-grams are applied to capture multi-word phrases (Lebret and Collobert, 2014) and it was proved to contribute to positive results in sentiment analysis (Dey, Jenamani and Thakkar, 2018). N-grams can alternatively be conceptualised as fundamental statistical models for language. N-grams are widely acknowledged as significant tools in several applications, including machine translation, spell checking, and speech recognition (Hirsimaki, Pylkkonen and Kurimo, 2009).

2.3 Content analysis VS Text analytics

2.3.1 Content analysis definition

Within the field of quantitative research, content analysis has emerged as a widely used and rapidly evolving technique. The proliferation of digital media and the development of increasingly sophisticated computer programmes have ushered in a new era characterised by faster and more efficient systematic examination of communications.

Krippendorff (2013, p. 24): Content analysis is a research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use.

Fico, Lacy and Riffe (2008, p.118) : Quantitative content analysis is the systematic and replicable examination of symbols of communication, which have been assigned numeric values according to valid measurement rules, and the analysis of relationships involving those values using statistical methods, to describe the

communication, draw inferences about its meaning, or infer from the communication to its context, both of production and consumption.

The definition of content analysis revolves around the fundamental operations that shape the characteristics of the data produced by this approach. The basis of content, including written materials, speeches, images, art, websites, and cultural objects, has its own distinctive features. The significance of these content extends beyond the researchers' understanding and reaches a wider audience. Researchers motivate themselves to recognise and comprehend the significance of comprehension, which is the main reason why content analysis is preferred over other investigative methods. Therefore, it is important to recognise that all texts, regardless of their physical format, are created and consumed by people who expect them to convey meaning beyond their physical form. Essentially, those who are skilled in language are able to go beyond the physical appearance of text and comprehend the underlying meaning within its structure.

Content analysis could be considered as the objective deconstruction of the features contained in messages. As a systematic and comprehensive framework, it could be conducted by human coders and computer-aided text analysis (CATA). Content analysis has a wide range of applications including evaluating human interactions in person-to-person contexts or character depictions from literary works and internet videos. It can be also applied to analysis the specific vocabulary usage in the textual data. Content analysis has found extensive application across a diverse range of research domains. This methodology has been effectively used. For example, content analysis has been utilised in studies with highly specific focuses, such as investigations into the queries posed by patients and their companions during physician-patient interactions (Eggly et al., 2006), studies on website engagement and Google Group thread structures related to both deceased and living public intellectuals (Danowski and Park, 2009), evaluations of emotional tone present in social media comments, often referred to as sentiment analysis (Thelwall, Wilkinson and American, 2009). Integration of content analytical metrics with other forms of measurement is a common practice.

Pian, Khoo and Chang (2014) exemplified this by conducting a study on user engagement in an online health discussion forum. Their strategy entailed using an eye-tracking system to identify text segments that captured users' attention by

recording their eye fixations. They then used content analysis to identify the categories of information that attracted the most interest. Similar research was carried out by Himelboim, McCreery and Smith (2013) who integrated network analysis and content analysis techniques to examine exposure to diverse political viewpoints on Twitter. They mapped ten politically contentious topic-based Twitter networks and identified communities of closely connected individuals. The investigation also analysed the political orientations of the message content. This investigation has demonstrated that Twitter users rarely encountered cross-ideological content within the clusters they followed, as the content within these clusters was noticeably homogeneous.

Moreover, it is customary to integrate content-analytic data with broader datasets, such as survey responses or experimental observations concerning message sources and recipients. By combing content-based analyses with contextual information, this method allows researchers to uncover a variety of insights. The main goal of quantitative analysis is to produce numerical data that represents the frequencies of significant categories and the measurements of other variables. Whether dealing with counts or quantities, the focus is consistently on a numerical approach. The main aim of quantitative content analysis is to obtain a numerical summary from a subset of chosen messages. This approach deviates significantly from creating all-encompassing impressions and providing comprehensive, elaborate explanations of individual messages or complete message collections. Instead, it prioritises quantifying and consolidating crucial message characteristics for analytical purposes.

It is possible to differentiate between the quantitative or qualitative nature of both the analysis and the investigated phenomenon. The central point under examination, however, remains of a qualitative nature. The properties of a message can be quantified on a regular basis (Smith, 2017), regardless of whether a study is labelled as "qualitative" but may also have significant quantitative dimensions.

2.3.2 From content analysis to text mining

In recent years, there has been an increase in the usage of terms to describe the numerous alternatives for analysing the content with computer algorithms. Computers have been instructed to obtain and process messages for diverse objectives, thus

greatly improving the potential of computational analysis applied in content analysis. Farrell, Wallis and Evans (2007) conducted individual and focus group interviews to examine attitudes towards nursing programmes. The researchers analysed qualitative data using standard codebook and content analysis techniques. The researchers analysed qualitative comments by categorising them into 23 distinct categories and themes. They then decomposed the data according to their occurrence frequency. In both instances, quantitative analytic methods were used to examine what was considered to be qualitative data.

Content analysis is a broad term that encompasses virtually any analytical technique designed to extract latent meaning from information. Content analysis includes both human and automated techniques. However, increasing computational power has enabled automated techniques to rapidly process immense amounts of data, thereby expanding the size and scope of datasets that can be analysed. Computational tools are now able to identify sentimental trends in speeches and map actor connections in social networks, etc. Content analysis is an umbrella term for a variety of analytical techniques designed to extract new insights from existing content.

Nowadays, in data analytics field, "data mining" comprises computational methods that facilitate the identification of complex patterns within big datasets, commonly referred to as "big data." Unlike conventional top-down techniques in statistical analysis, data mining implements algorithms that iteratively construct patterns, progressively increasing clarity with each case processed Nisbet, Elder and nd Miner (2009). A prominent instance of data mining can be observed in the field of social media, where the aim is to examine the online behaviour of individuals from a business-oriented perspective.

When working with textual data, the term 'data mining' is replaced with 'text mining'. Text mining involves extracting meaning from sequences of characters, much like reading (Bholat et al., 2015). However, text mining differs from standard reading in two significant ways. Firstly, text mining procedures driven by computers can process and synthesise large quantities of text, which exceed the temporal limitations of human reading. Secondly, these computational methods have the potential to extract meaning from text that humans often miss due to their tendency to ignore patterns that contradict their predetermined beliefs and expectations. Two main approaches arise in the field of text mining: unsupervised machine learning and

supervised machine learning. Unsupervised machine learning aims to uncover significant patterns concealed within unstructured textual data. On the other hand, supervised machine learning involves the iterative training of computer programmes via human classification of a textual corpus.

Natural language processing (NLP) is a more comprehensive approach to the study of large textual datasets. Essentially, NLP entails utilising computational systems to decode and manage natural language. The efforts made by computer scientists to educate algorithms in comprehending human language resulted in the emergence of natural language processing (NLP), thereby colliding with the domain of artificial intelligence (AI). Early goals in the field of NLP aimed to integrate speech recognition capabilities into natural language processing systems, enabling them to comprehend human speech (Bates and Weischedel, 1993). Moreover, the field of Natural Language Processing (NLP) has adopted pragmatic and utilitarian goals. These objectives encompass various undertakings, such as language translation, data mining for question answering, and enabling human-machine interactions to offer guidance. Similarly, CATA integrate various functionalities previously conducted manually through laborious coding efforts. This prompts the relevant topic of comparing automated techniques against the traditional practice of hand-coding. Both text mining and CATA used the computer as a tool and have significant advantages over human-centred methods, particularly in three essential domains:

Reliability. Ensuring the consistency of results poses a substantial challenge in research attempts that use large teams of human coders. Moreover, it is possible for the programmer to accidentally modify their coding standards throughout the duration of the project, which could be possibly affected by the content they are coding. To address these problems and ensure an essential degree of consistency in the findings, it is important to conduct a range of evaluations, in which the purpose of is to maintain a consistent standard of outcomes and offer an unbiased measure of accuracy. However, computer-based techniques perform without being subject to these limitations. The results can be systematically measured, thereby clarifying the rationale underlying the machine's choice of a particular response.

Reproducibility. The problem of reproducibility presents an important disparity between the establishment of coding standards and the implementation by individual coders. Despite attempts to establish measures for dependability and

consistency, it is possible that the repetition of a project by several groups of programmers might produce marginally differing findings. On the other hand, a computer-implemented coding system that complies to a defined set of rules would consistently provide similar outcomes, even with a high number of iterations.

Scale. Furthermore, the issue of scale presents a substantial challenge in which the utilisation of limited text samples need to be analysed to provide quantifiable outcomes within a period. Computational techniques offer a significant benefit in terms of their speed and ability to be scaled up. One computer has the potential to go above the productivity of numerous human coders, as it can operate continuously without requiring any breaks. Furthermore, the process of increasing capacity largely involves the incorporation of an additional software or other computing resources typically at a relatively low expense, allowing continued operation for a period spanning many years.

2.3.3 Syntax versus semantics

The inherent structure in textual data is influenced by the grammatical structures of natural language. For example, an academic article contains essential elements such as a title, list of authors, abstract, keywords and introduction. Each of these elements consists of a collection of words, symbols, phrases and sentences, which are arranged into paragraphs, sections or subsections.

In the context of each sentence, there is a combination of words from many linguistic categories, including nouns, verbs and adjectives. They could be integrated into verb-noun or prepositional phrases to effectively communicate substantial concepts. These terms which are primarily from the vocabulary of their respective natural languages could be brought together through the use of linguistic principles known as *grammar*. Various languages have distinct sets of grammatical rules that enable the production of diverse written compositions within each language. The principal objective of text mining algorithms is to enable the examination of textual data without being dependent on human comprehension. Nevertheless, the scope of computer analysis is restricted to the direct examination of individual letters inside words and the arrangement of these words. Instead of an in-depth understanding of the transmitted information, the capabilities of the computer are restricted to the analysis of the structural and syntactical elements of the text.

Syntax can be considered as structure of words and how these words are the arranged systematically in sentences and paragraphs based on the established language norms and grammatical rules. Based on these adherences to the rules, there are various statistical patterns that can be often found in large bodies of text. Therefore, text mining can utilise some computational tools to discover the syntax within text because of the presence of a consistent structure. However, it is important to recognise that the study of syntax alone is not sufficient to achieve a comprehensive understanding of meaning. In contrast, the field of *semantics* investigates the significance of specific lexical units within the corresponding contextual settings. The attempt of discovering the comprehensive semantic importance of a given text is a challenging task, which requires an extensive understanding of the deployed language (Wachsmuth, 2015).

Nevertheless, the examination of syntax on its own possesses practical applicability within the field of text mining, even in the absence of semantics. The occurrence of this phenomena can be attributed to the complex interplay between syntax and semantics. Numerous text mining subcategories, such as document categorization and information retrieval, concentrate on the prioritisation or retrieval of particular document categories within large datasets. The algorithms discussed in this context are grounded on the fundamental premise that the presence of similar words, referred to as syntactic congruence, signifies a common underlying meaning, referred to as semantic similarity. The success of these approaches is demonstrated by their reliance on syntactic clues because publications that share a large number of common terms frequently relate to related subjects. On the other hand, in certain situations, the focus shifts to the study of meaning and interpretation. Concept extraction involves the automatic recognition of words and phrases that have the same meaning. In these specific cases, text mining approaches need to use grammar as a means of inferring the underlying semantic links. A prominent example that supports the integration of multiple components is the Generalised Vector Space model.

2.4 Research gaps

We have reviewed the literature related with unstructured data, its application in the organisations, features of text as well as the content analysis framework. As such, there are several research gaps which could be filled by this thesis.

First, the complicated structure of textual data, resulting from linguistic complexities, presents substantial difficulties for text mining. Further exploration is required in the domains of advanced semantic analysis techniques, such as context-aware computing, which involves algorithms capable of comprehending the context in which words are employed. Texts do not possess intrinsic singular meanings that can be definitively ascertained or fully described in isolation or directly linked to their sources. Understandings towards textual data can range from quantitative evaluations, such as counting characters, words, or sentences, to categorising expressions, analysing metaphorical elements and explaining logical compositional structures. Therefore, the development of more advanced models that integrate linguistic and context frameworks is needed to improve the accuracy of text mining tools in interpreting text.

In addition, compared with content analysis, the application of text mining does not have a standardised methodological framework, especially when moving from human to automated or semi-automated procedures. The presence of this gap has a negative impact on the dependability and accuracy of research findings as many studies may employ multiform text mining flows. This thesis could explore the procedures and benchmarks for text mining which could facilitate the attainment of enhanced uniformity. This thesis could exploration the creation of benchmark and validation measures for diverse domains in order to establish relatively standardised text analysis approaches across many disciplines of study.

What's more, the scalability and reproducibility of text mining techniques are crucial as the amount of textual data expands rapidly. Existing tools and approaches may lack efficiency in processing large-scale data sets or providing quick analytics for large textual datasets. This thesis could contribute to creating novel algorithms and data processing frameworks that improve computing efficiency. Besides, the discrepancy exists between the formulation and implementation when people are involved in the coding procedure. In contrast, texting mining techniques could consistently comparable results even for large textual datasets.

CHAPTER 3 Incorporating topic membership in rating prediction from unstructured data: a gradient boosting approach

3.1 Introduction

Online reviews are a longstanding topic in literature, and their influence on sales has been well documented (See-To and Ngai, 2018; Liu, Lee and Srinivasan, 2019; Zhang, Tian and Fan, 2020). Customers view reviews as a way to gather information about the quality of products or services (Li, Wu and Mai, 2019; Zhao et al., 2019) as well as embodiments of experience-specific information regarding the quality of products or services. The rating score from customer feedback represents customer overall satisfaction or dissatisfaction directly thus affecting customers' purchase behaviour. As the antecedents of customer rating have been extensively researched, the problem of rating prediction attracts more attention, especially when trying to attribute the rating score to particular aspects of the review or service also known as aspect rating (Korfiatis et al., 2019a). The textual content associated with the rating score is provided to justify the latter, thus explaining the underlying rationale, which offers consumers opportunities to express themselves freely rather than feel restricted by pre-defined subareas defined in the user interface (Büschken and Allenby, 2016). Compared with using the characteristics of reviews (i.e., review length) for rating prediction, research concerning the prediction of rating scores from the textual content of customer reviews is attracting more attention, as it tends to carry significant implications for particular service domains.

The review rating score prediction problem is considered to originate from the sentiment classification task of classifying reviews as thumbs up (recommended) or thumbs down (not recommended) (Pang, Lee and Vaithyanathan, 2002; Qiu et al., 2018). Several studies suggest a high level of consistency between a review's rating score and its textual justification (Qiu et al., 2018; Hu, Koh and Reddy, 2014). Even the significant variation of ratings could be explained by customer sentiment statistically (Geetha, Singha and Sinha, 2017), the numeric rating scores in customer reviews do not equally represent customer sentiment, as the polarity information appearing in reviews cannot be fully captured by ratings due to the limitations of the rating scale itself (Ghose and Ipeirotis, 2011).

Online reviews are widely adopted for companies to understand how customer perceived the quality of products or services. As demonstrated by Tirunillai and Tellis (2014), multidimensionality exists in quality measurement. Parasuraman et al. (1988) developed SERVQUAL to measure service quality using multiple dimensions including reliability, tangibles, responsiveness, assurance, and empathy. Similarly, there are also latent dimensions within the textual content in online reviews. The predictive power of these dimensions has been examined by several studies. The extracted dimensions in review text could add predictive accuracy to the overall customer satisfaction (Korfiatis et al., 2019a). Xu (2020) also revealed the textual factors in review text have an asymmetric influence on customer satisfaction.

The latent dimensions can be discovered using topic models, which has been applied in various business areas to identify different themes from customers' textual feedback. Using topic models, the association and connection between a rating score and review text can be established, which in turn can assist businesses in understanding the reasons driving different levels of customer satisfaction as well as the multidimensionality of the service's outcome (Korfiatis et al., 2019a).

Among various types of topic models, LDA is considered as the appropriate method from two aspects. First, there are two types of topic models including single-membership topic models and mixed-membership topic models. Single-membership allocate each document to only one topic while mixed-membership models indicate that a document is considered as a mixture of multiple topics, in which each word belongs to one single topic. Mixed-membership models allow each document cover multiple topics. LDA allow us to extract the latent topics (dimensions) as well as how much each document is associated with each latent topic, which is the topic membership. Topic membership contained in the unstructured data could increase the accuracy when predicting rating scores (Tirunillai and Tellis, 2014).

Second, LDA is considered as a milestone among the development of topic models. Compared with the earliest topic models (e.g, LSA) which reduces the document dimensionality by applying a singular value decomposition (SVD), LDA is based on the Bayesian statistical topic modelling method which not only improves the model performance but also provides clarity for interpretability issues. Based on LDA, more and more topic models are developed, such as CTM and STM. Ideas of these models are similar with LDA and are considered as the extension of LDA.

Considering the main objective of study, we selected LDA as the topic modelling method as it is the most widely employed and acts as the foundation of other extensions.

Therefore, in this study, we aim to evaluate the ability of topic membership to explain and predict customer overall satisfaction as well as exploring how topic membership could be coupled with customer sentiment to enhance the prediction accuracy. Specifically, using machine learning, we examine *how topic models can be used in conjunction with sentiment analysis to quantify customer feedback in cases where domain specific sentiment dictionaries cannot be available.* In addition, given the limitations and the issues with domain-specific sentiment analysis, we evaluate whether *topic modelling can provide a better alternative than sentiment analysis* in review rating prediction. The latter has significant implications, as it removes the necessity for employing domain-specific dictionaries and other approaches that need to be adjusted to the vocabulary of the business domain.

To address these objectives, we perform a large-scale machine learning analysis on a dataset of 1,810,831 customer reviews from 12,153 restaurants on *JustEat*, a popular online food delivery platform in the United Kingdom. Our analysis is multi-faceted. We design and validate two experiments. First, a binary classification task is formed (using a rating score cut-off) to have a robust evaluation of the performance of topic membership as well as its comparison with customer sentiment (in both document-level and sentence-level). We also compare the predictive abilities across topic memberships of all topics. Second, the task of predicting the actual rating score is proposed. We combine each topic membership separately with the polarity score and compare the additive predictability of each topic in predicting the rating score. We perform a post-hoc analysis for robustness by incorporating these two features as covariates using a gradient boosting model using XGBoost—an established machine learning technique. Using the corresponding Shapley additive explanation values (SHAP), we identify the contribution to the prediction of the rating value by each topic/sentiment combination separately.

Our study contributes to the analytics literature by demonstrating how these two different approaches of incorporating features from unstructured data can be used in tandem to predict and explain customer satisfaction. Our findings can lead to faster and more accurate managerial insights for businesses based on the characteristics of

online reviews and the intrinsic shortage of sentiment analysis. On one side, the large volume of online reviews increased the difficulties in interpreting and comprehending the meanings of more consumers. However, the topic-based rating prediction can directly extract the detailed dimensions that consumers complain or praise about and also link them with the quantitative information (review score), which contribute to the faster decision-making process. On the other side, the power of sentiment analysis could be affected because of the existence of two-sided reviews which cover both positive and negative aspects while topic-based rating prediction can uncover multidimensional aspects of service quality that cannot be captured by sentiment analysis.

Thus, it can explain customer rating scores and highlight service aspects within customer textual comments to facilitate businesses' understanding of customers' perceived quality towards products or services, which affects customers' purchase decisions (Yeo et al., 2022). Beyond customer reviews from review websites, there are also various sources of customer feedback used in business analytics cases, such as online forums and social media, which do not contain rating scores, or the rating score is incomplete for some of these dimensions. Our study can also extend to these sources.

To this end, the rest of this paper is organised as follows: Section 2 reviews the literature on sentiment analysis and topic modelling on customer reviews and how they are applied for rating prediction. Section 3 demonstrates our data sample, models, and metrics for model performance evaluation. Section 4 details the corpus pre-processing procedures and model parameter selection; then, it displays the results for these experiments. Section 5 summarises the analysis and discusses the implications of the results for researchers and practitioners. The study concludes in Section 6 with limitations and future research directions.

3.2 Literature review

3.2.1 Sentiment analysis

Sentiment analysis is a celebrated computational analysis method in business and management used for detecting customer attitudes and feelings expressed within an unstructured part of customer feedback through natural language processing (NLP) techniques. It is comprised of two main tasks: (a) polarity detection—the

identification of whether the text is positive or negative and (b) affect detection—the feelings and emotions expressed in the written communication. Polarity detection is the most common approach applied in sentiment analysis, mainly thanks to the ease of labelling the large textual corpus concerning consumer interaction with company touchpoints (e.g., via social media). In that respect, the literature treats polarity as either an ordered categorical (with the text classified as positive/neutral/negative) or a continuous numerical score of an asymmetric continuum ranging from a normalised negative algebraic value to a positive algebraic value (e.g., -1 to +1). Depending on the task at hand, sentiment can be calculated at the document or sentence levels. The latter also produces a confidence band that can produce more reliable prediction outcomes if sentiment is operationalised as input.

Apart from document-level and sentence-level sentiment analysis, aspect-level sentiment analysis is also discussed in the past literature. In many situations, it requires more investigation at the aspect level to identify entities and the associated aspects and sentiments. For instance, companies would like to identify what aspects of their products attract or dissatisfy customers from customer reviews to improve products (Birjali, Kasri and Beni-Hssane, 2021). There could be two types of aspects: explicit aspects and implicit aspects. The former represents those aspects that are directly mentioned in the text, the latter illustrates those aspect terms that don't appear in the text but are implied by other terms (Hu and Liu, 2004). Several Machine learning approaches has been employed to extract the implicit aspects (Bagheri, Saraee and De Jong, 2013; Quan and Ren, 2014).

Various sentiment analysis techniques are discussed in the literature (Liu, 2010, 2012; Al-Natour and Turetken, 2020; Yadav and Vishwakarma, 2020). These can be summarised into two major types: lexicon-based and machine learning approaches. The lexicon-based sentiment approach uses a bag-of-words model that requires a dictionary consisting of predefined words or phrases assigned by negative or positive values. The other popular approach considers sentiment analysis as a pattern recognition problem and utilises machine learning techniques for classification tasks or predictions (Ghiassi and Lee, 2018). Farkhod et al. (2021) proposed a TDS (Topic Document Sentiment) model, which is an unsupervised machine learning method based on the JST (Joint Sentiment Topic) model and LDA. They used the proposed model to discover the sentiment at the word, document and

topic levels. When compared with sentiment terms (usually adjectives) from the lexicon-based approach; machine learning techniques can extract broader and more comprehensive features about several aspects of text, including nouns and verbs expressing descriptions and attitudes towards these objects (Liu, 2012).

In addition, the hybrid approach which combines lexicon-based approach and machine learning approach attracts more attention as it can integrate the advantages of machine learning approach (high accuracy and flexibility) and that from lexicon-based approach (stability) (Birjali, Kasri and Beni-Hssane, 2021). For instance, Marshan, Kansouzidou and Ioannou (2020) developed a hybrid model combing the lexicon-based and machine learning approaches to detect customer sentiment contained in reviews from an e-commerce platform. Three machine learning approaches are selected including the Naïve Bayes, KNN (k-nearest neighbours) and SVM (Support Vector Machine). Results showed that the Naïve Bayes had the best performance as a classifier. Elshakankery and Ahmed (2019) proposed a hybrid model HILATSA, representing Hybrid Incremental Learning approach for Arabic Tweets Sentiment Analysis to detect the sentiment in tweets. It is a semi-automatic learning system which will update the lexicon to keep it up to date.

Sentiment analysis for detecting customers' emotions and attitudes, has been widely applied in user-generated content (e.g., online reviews), and several studies have examined the importance of customer sentiment in understanding the relationship with customer ratings, predicting sales, and identifying fraudulent reviews (Zhao, Xu and Wang, 2019; Zhao et al., 2019; Kumar et al., 2022). Innovative applications, combining machine learning and bag-of-words-based approaches, have been applied in practice. Dey, Jenamani and Thakkar (2018) proposed a system that could generate Senti-N-Gram, an n-gram sentiment dictionary, and proposed an algorithm to extract the sentiment scores for n-grams from a random corpus consisting of review text as well as numerical ratings. This approach showed better performance than an existing unigram-based approach (VADER) and another n-gram-based approach (SO-CAL). Recent studies have also applied machine learning approaches to expand dictionary coverage. For instance, Sharma and Dutta (2021)proposed a framework called *SentiDraw*, which calculates the sentiment score for each word from customer reviews based on the rating distribution. Then it was combined with Support Vector Machine (SVM) to achieve better polarity determination.

3.2.2 Topic models on user-generated content

Online reviews, functioning as the *"voice of the consumer"*, are a form of electronic word-of-mouth (eWOM); these play a critical role in affecting customers' decision-making process, behavioural intention, and product sales performance (Li, Wu and Mai, 2019; Verma and Yadav, 2021). It has been considered an unignorable information source for both customers and sellers, especially the textual content within customer reviews, which includes the textual description of the first-hand usage experiences of previous customers (Guo, Barnes and Jia, 2017). The growing popularity of online reviews provides customers with the opportunity to express themselves naturally with unstructured data.

Customers' opinions in online reviews are multidimensional and may reflect different aspects, such as product-specific features or service aspect related evaluations (Büschken and Allenby, 2016; Kim, Park and Lee, 2020; Mai and Le, 2021). Therefore, to ascertain these latent dimensions from customer reviews, the topic modelling approach is applied widely, as topic models could discover patterns reflecting latent topics within a document from unstructured customer reviews. It assumes that documents consist of a set of topics, and each topic covers a mixture of words (Alghamdi and Alfalqi, 2015). There are a variety of topic models, including Latent Semantic Analysis (LSA) (Landauer, Foltz and Laham, 1998), PLSA (Hofmann, 2001) and LDA(Blei, Ng and Jordan, 2003). There are many extensions of LDA, including the Correlated Topic Model (CTM) (Blei and Lafferty, 2007) and the Structural Topic Model (STM) (Roberts et al., 2014).

Researchers either adopts existing topic models or proposed new variants of topic models to discover the multidimensionality of customer reviews. The latent dimensions contained in customer reviews are critical since they serve as the foundation for how customers evaluate service, brands and firms, thus affecting new product development or brand positioning. Kwon et al. (2021) employed topic modelling approach and sentiment analysis to online customer review for airlines in order to identify customers' needs. They extracted six dimensions using LDA and identified several words that contained positive and negative emotions respectively.

Tirunillai and Tellis (2014) extended LDA and employed the variant of LDA to customers reviews from five markets and 16 brands. They identified the latent

dimensions and ascertained the valence, dynamics and heterogeneity, etc. for strategy analysis. Büschken and Allenby (2016) developed a new model (sentence-constrained LDA model) based on LDA. They believe that people tend to change their topics across sentences instead of discussing two topics in one sentence. They applied it to two datasets consisting of customer feedback from both restaurant and hotel industry, illustrating the helpfulness of topic modelling approach for unstructured data. STM was also employed to customer reviews from hotel industry in order to understand customer dissatisfaction from their complaints (Hu et al., 2019). They identified top 10 latent dimensions related with customer dissatisfaction and how dimensions change across hotel grades. Customers of high-grade hotels mainly complained about service issues while that of low-graded hotels are more dissatisfied with facility-related problems.

3.2.3 Rating prediction

Given the importance of understanding customer satisfaction, rating prediction is a vital task. Several studies have examined the characteristics of online reviews (e.g., review length) to understand customer ratings (Ghasemaghaei et al., 2018; Lai, Wang and Wang, 2021). The information contained in textual comments also plays an important role in understanding customer ratings. Therefore, how to utilise review texts to predict customer rating scores has been a popular topic.

Based on whether the review text is provided in the prediction, the rating prediction task is mainly divided into two categories: (a) personalised rating prediction and (b) review-aware rating prediction. The first focuses on predicting users' rating scores over unrated items using their previous rating behaviours, which is widely explored in the recommendation system field (Zhang and Wang, 2016; Cheng et al., 2018). Latent factor models, including matrix factorisation, are applied widely and successfully for this type of rating prediction. Customers' textual comments could be corroborated to model user interests and item features (Tan et al., 2016) and to improve the accuracy of rating prediction models. The second concentrates on understanding customers' rating scores by discovering valuable information from the provided review text. The direct relationship between sentiment and ratings has been confirmed in previous studies (Geetha, Singha and Sinha, 2017).

Table 3 summarises the two categories of rating prediction tasks, which approaches they adopt to extract features from the review text, and what prediction models they adopted.

Table 3 Overview of existing literature in rating prediction

Type	Task	Approach	Description	Prediction model	Indicative Studies
Personalised rating prediction	Extracting contextual features and combining them to rating prediction	Topic model extensions	*These studies were based on topic models, and combined topic modelling approaches with the latent factor models.*	Self-proposed model	(McAuley and Leskovec, 2013; Zhang and Wang, 2016; Cheng et al., 2018)
	Extracting complex features and improving rating accuracy	Deep learning	*These studies built deep learning architectures with multiple layers.*	Self-proposed model	(Seo et al., 2017; Xing et al., 2019)
Review-aware rating prediction	Exploratory relationship with rating	Sentiment analysis	*Direct impact of sentiment on review ratings and obtained the preliminary relationship between sentiment and ratings.*	Fix effects model/linear regression	(Hu et al., 2014; Geetha et al., 2017)
	Rating prediction	Sentiment analysis	*Applied sentiment analysis to review text and adopted the extracted sentiment features for rating prediction.*	Ridge regression/linear regression	(Qu et al., 2010; Y. Zhao et al., 2019)
	Rating prediction	Topic models	*Applied LSA to discover positive and negative attributes from reviews*	Text regression	(Xu, 2020)

			and used these attributes as independent variables in text regression to predict overall satisfaction.		
	Rating prediction	Both sentiment analysis and topic models	*Applied LDA to exploit several dimensions from review text and combined them with sentiment detected by sentiment analysis to explain the overall satisfaction.*	Linear regression/multinomial regression	(Xiang et al., 2017)
	Rating prediction	Self-proposed model	*Added constraints to the LDA model and proposed a sentence-constrained LDA model, and combined it with rating data.*	Self-proposed model	(Büschken and Allenby, 2016)

A variant of LDA was employed to extract latent topics and to predict customer ratings using a dataset of customer reviews from Italian restaurants. It is examined that topic membership could be a meaning device to explain customers' ratings scores. They used a latent cut-point model to examine the relationships between customer satisfaction and the topic membership of 8 topics (Büschken and Allenby, 2016).

LSI was also applied to the content of customer reviews from hotel industry and identified 8 positive factors and 17 negative factors. Text regression was conducted using the vector space of each textual reviews to examine how they can affect the overall customer satisfaction. The asymmetric effects were found from the results representing that not all positive textual factors affected customer satisfaction positively (Xu, 2020). STM was also adopted to online reviews from airline passengers and extract latent dimensions of service quality from the textual content. Together with the predefined subcategories by the online platforms, the latent dimensions could add the ability to predict customer overall satisfaction (Korfiatis et al., 2019b). These studies provide us with another way to predict the customer ratings in addition to using the predefined subscales by companies or platforms and proved the ability of topic membership to predict and explain customer ratings.

3.2.4 Research gaps

Among the two types of rating prediction, our work is related to the second stream where the review text is given and could be utilised in review-aware rating predictions. Instead of focusing on a different computationally complex approach, our goal is to generate new features that can better incorporate the information in the review text. For that, we select two unsupervised text mining approaches used in previous studies to exploit customers' emotions and opinions. We compare this study with previous studies which focus on the review-aware rating prediction as shown in below table.

Table 4 A comparison with other studies

Studies	**Approach**		**Prediction model**	
	Sentiment analysis	*Topic models*	*Regression*	*Machine learning*
Hu et al. (2014)	✓		✓	
Geetha et al. (2017)	✓		✓	
Qu et al. (2010)	✓		✓	
Y. Zhao et al. (2019)	✓		✓	
Xu (2020)		✓	✓	
Büschken and Allenby (2016)		✓		✓
Xiang et al. (2017)	✓	✓	✓	
This study	✓	✓		✓

These studies provide us with two ways to predict the customer ratings proved the ability of topic membership and sentiment to predict and explain customer ratings. Compared with those studies that only adopt one approach in mining features from review text, our approach examines and compares the adoption of two approaches in predicting customer rating scores. In addition, we also combine the two approaches for evaluating the usefulness of topic modelling to the predictive power of sentiment analysis. As to the prediction model, instead of linear or logit regression models, we adopt machine learning algorithms for the rating prediction as well as comparing the abilities of different machine learning algorithms, which fills the research gaps in a methodological aspect.

3.3 Data and methods

For this study we follow the Cross-Industry Standard Process for Data Mining (CRISP-DM), which is proposed by Wirth and Hipp (2000), aiming to converting business problems into data mining projects which could be carried out and applied

regardless of the type of technologies and industries. We illustrate "Business Understanding" and "Data Understanding" in Sections 3.1 and Section 3.2. We move to "Data Preparation" in this section and "Modelling" in Section 4.

3.3.1 Dataset

Our data considers textual reviews of UK customers and is collected from *JustEat*, the most popular online food delivery service provider in the UK, with more than 68% of the market share of online orders[1]. Besides, *JustEat* could provide customers with each review including rating score and review text from previous customers who purchased in the specific restaurant while other competitors (e.g., *Deliveroo* and *UberEats*) only display the overall rating score and the number of reviews of the specific restaurant. After customers order and receive the delivered food, they are invited to leave customer reviews describing the entire experience with the food delivery. Potential customers searching for a restaurant can find these reviews on restaurants' pages, which can be of assistance to them. We collect customer reviews written in English and published on their website from January 2016 to November 2021. Generally, there should be a numerical rating score with textual justification in each review. Figure 5 shows an example of one customer review for a restaurant. *JustEat* adopts a 6-point scale rating system in which customers give a rating from 1 to 6 stars for three service categories: food quality, delivery time, and restaurant service. The final rating score shown to other consumers when reading these reviews is calculated as the (simple) average of the individual ratings of these three categories.

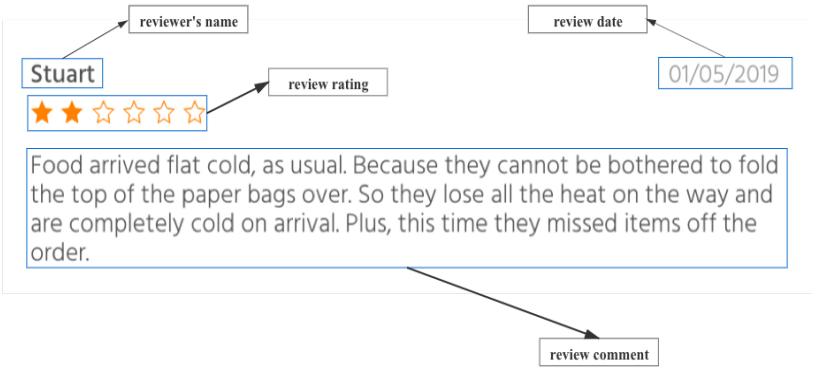

Figure 5 An example of customer reviews on JustEat

[1] Statista Global Consumer Survey – Brand Report, 2021

The textual justification provided by the customer considers all three service categories; therefore, the average rating provides intervals between the minimum and maximum rating scores. To make our analysis more meaningful, we select reviews with textual comments from customers and filter out those that only contain rating scores. Additionally, each review length is constrained in terms of length between 15 and 200 words[2]. In total, our sample contains 1,810,831 customer reviews from 12,153 restaurants. In this paper, unlike many other datasets in which reviews are rated given a 5-point rating system (Hu et al., 2014; Geetha et al., 2017), the reviews from JustEat adopt a 6-point scale rating system in which customers give a rating from 1 to 6 stars for three service categories: food quality, delivery time, and restaurant service. The final rating score shown to other consumers when reading these reviews is calculated as the (simple) average of the individual ratings of these three categories, leading to 16 groups (i.e., 1.0, 1.3, 1.7, …, 6.0)

In the study from Pang, Lee and Vaithyanathan (2002), the reviews are classified into three categories: positive, negative and neutral according to the rating. They excluded the neutral category as their concentrated only on discriminating between positive and negative sentiment. As we aim to conduct a robust check focusing on the classification of positive and negative classes, we used the median of the rating scale (3.5 stars) as the boundary (marked with the blue line in Figure 6) for separating the positive and negative reviews given its even distribution in both classes as shown in the density distribution of the ratings as provided in the dataset.

[2] A winsorization procedure was followed for the maximum values considering that any reviews about 200 words were above the 95% quantile of the distribution of review word length.

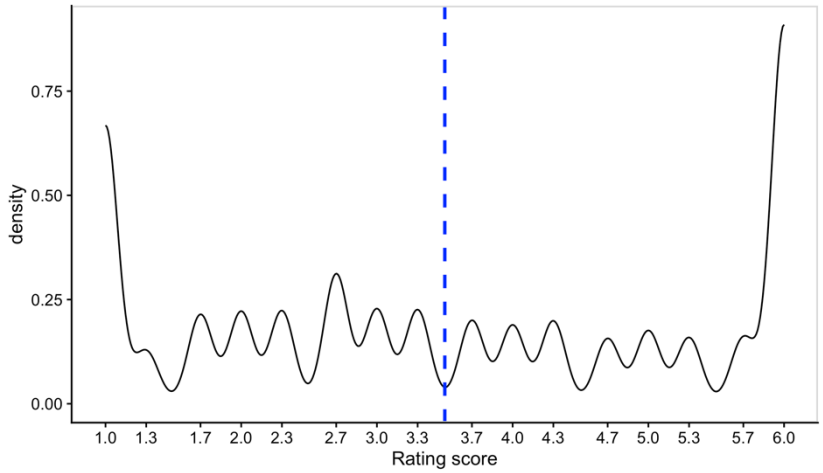

Figure 6 Distribution of rating scores within the reviews in our sample. The blue dashed line outlines the median rating score.

Figure 6 shows the distribution of the rating scores for our sample. This indicates that the percentages of extreme ratings are significantly higher than others. The average rating score is M=3.57 ($SD = 1.82$), which is slightly positive. As to the actual distribution of the scores, the highest percentage occurs in 6-stars ratings, amounting to 20.9%, followed by a 1-star rating, which has the second-highest proportion (15.3%). In total, the proportions of positive (49.19%) and negative reviews (50.81%) in our sample are similar, which means that our sample is balanced.

3.3.2 Sentiment analysis

Sentiment analysis is generally measured through polarity, which measures the positive/negative intent expressed in the text. It is generally calculated by various techniques related to the bag-of-words approach; however, other studies in the literature have also applied more complicated models. For polarity calculation, the standard method is to use a lexicon-based approach with domain-specific or general dictionaries (or lexicons). These lexicons can be compiled manually or acquired automatically. The function we adopt to calculate polarity is based on subjectivity lexicons, which contain a list of terms connected with particular emotional states. For instance, the word ‘awful’ relates to a negative state, while the word ‘excellent’ is associated with a positive state.

We employ the subjectivity lexicon from Hu and Liu (2004) to calculate polarity, including approximately 6,800 prior-labelled words, which have been identified by benchmarking these terms on a large dataset of online consumer reviews.

For identifying polarity in the text, a cluster of terms (x_i^T) that contains four words before and two words after the polarised word has been used to introduce context. Words in the cluster that do not have value are called neutral words and are tagged as x_i^0, which affect word count (n). In addition, there are words that do not have emotion but have an influence on the emotional context, such as valence shifters. Amplifiers (x_i^a)/De-amplifiers (x_i^d)/ are words which can increase/decrease the emotional intent of words, which are given a weight for calculation (Rinker, 2020). The context is defined as follows:

$$x_i^T = \sum((1 + c(x_i^A - x_i^D)) * w(-1) \sum x_i^N \quad (1)$$

where:

$$x_i^A = \sum(w_{neg} * x_i^a), x_i^D = \max(x_i^{D'}, -1) \quad (2)$$

$$x_i^{D'} = \sum(-w_{neg} * x_i^a + x_i^d) \quad (3)$$

$$w_{neg} = \left(\sum x_i^N\right) mod2 \quad (4)$$

The polarity score is calculated as:

$$\delta = \frac{x_i^T}{\sqrt{n}} \quad (5)$$

We constrain the polarity score at (-1,1) using a transformation formula as follows:

$$\left[\left(1 - \frac{1}{\exp(\delta)}\right) * 2\right] - 1 \quad (6)$$

We use consumers' original review comments before Part-of-Speech tagging and stop words removal because the absence of words will decrease the accuracy by affecting the density of keywords.

3.3.3 Latent Dirichlet Allocation (LDA)

The LDA model proposed by Blei et al. (2003) is an unsupervised learning model based on Bayesian inference. Its underlying principle is exchangeability. Compared

with latent semantic indexing (LSI) and probabilistic LSI (pLSI) models, LDA considers the exchangeability of both documents and words. LSI applys statistical computations to a large corpus of text to extract and represent the contextual usage meaning of words (Batra and Bawa, 2010).

LSI adapts Singular value decomposition (SVD) into the term-document matrix to achieve dimensionality reduction (Zelikovitz and Hirsh, 2001). It is demonstrated that several basic linguistic notions (e.g., synonymy and polysemy) could be captured by the linear combinations of the tf-idf features, which are derived by LSI (Deerwester et al., 1990). However, the biggest weakness of LSI is the lack of satisfactory statistical foundation. Subsequently, the probabilistic LSI (PLSI), with a solid statistical foundation using a probabilistic method replacing SVD, is proposed by to address the weakness of LSI (Hofmann, 2001). It considers each word as a sample from one mixture model, in which the mixture components (multinomial random variables) are considered as "topics". In pLSI, a list of mixing proportions for these mixture components is used to represent each document. However, no probabilistic model at the document-level is not provided, which mean the numbers from each list is not from any generative probabilistic model, leading to overfitting problems seriously (Blei et al., 2003).

LDA is a generative probabilistic model that can deal with sparse vectors of discrete data, including bag-of-words in text data and image features. For text data, the core assumption is that each document is considered a random mixture of latent topics, while each topic is represented by a multinominal distribution over words. LDA is based on the assumption that the author of each document would have the same probability to use same words when writing the same "*topic"*. LDA is a generative probabilistic model, which simulates the process of an author producing a document. In this process, the probability of writing a word is related with the topic that are written about. However, if two authors have different words to write for the same document, the distribution of words might change to an unrelated topic by making inaccurate inferences.

Every document is created by a list of hypothetical and unobservable *'topics'*. Each document is assumed to be presented by a mixture of topics that reflect distributions sharing common Dirichlet priors. In a single document, the probability of each topic is between 0 and 1, with the sum of them amounting to 1. The extent to

which documents are associated with topics is considered document-topic proportions, also known as topic membership. The latent topics are considered as the distributions over a fixed vocabulary in which each word has a possibility belonging to each latent topic.

In this study, each review is a single document. We index each review as $r \in (1,2 \dots, R)$. K presents the number of topics, which is the primary input variable. The generation process is summarised as follows:

- For each topic $k \in (1,2 \dots, K)$, draw a Dirichlet distribution over the vocabulary V, $\beta_k \sim Dir(\eta)$.
- For each review, r, choose $\theta_r \sim Dir(\alpha)$.
- For each word w_a, from review r, a topic assignment is drawn from a multinominal distribution over $\theta_r, z_{r,a} \sim Mult(\theta_r)$, where $z_{r,a}$ represents the word-specific topic assignment.
- The observed word, $w_{r,a}$ is drawn from $Mult(\beta_{z,r,a})$, where $w_{r,a} \in (1,2 \dots V)$.

The joint distribution of all unobserved variables and observed variables is expressed as follows:

$$p(\beta_K, \theta_R, z_R, w_R | \alpha, \eta) = \prod_{k=1}^{K} p(\beta_k|\eta) \prod_{r=1}^{R} p(\theta_r|\alpha) \prod_{a=1}^{N} p(z_{r,a}|\theta_r) p(w_{r,a}|z_{r,a}, \beta_{r,k}) \quad (7)$$

Figure 7 depicts the graphical model of the LDA process in plate notation. The shaded nodes present the only observed variables $w_{r,a}$, which represents the ath word in review r. A is the total number of words in each review. Each review could be represented as a mixture of topics. θ denotes the review-topic distribution, which indicates how much each review is related to topics. β denotes the per-review topic–word distributions. The problem of sparsity caused by the large vocabulary occurs in many document corpora. The *smooth* method is usually adopted to avoid assigning zero probability to words that are from new documents but don't appear in documents from training corpus (Blei et al., 2003). Instead of the commonly used method-Laplace smoothing, LDA places the Dirichlet priors on the multinomial parameters. Therefore, α and η, as two Dirichlet parameters, denote the smoothing of topics with reviews and words within topics, respectively (Syed and Spruit, 2017).

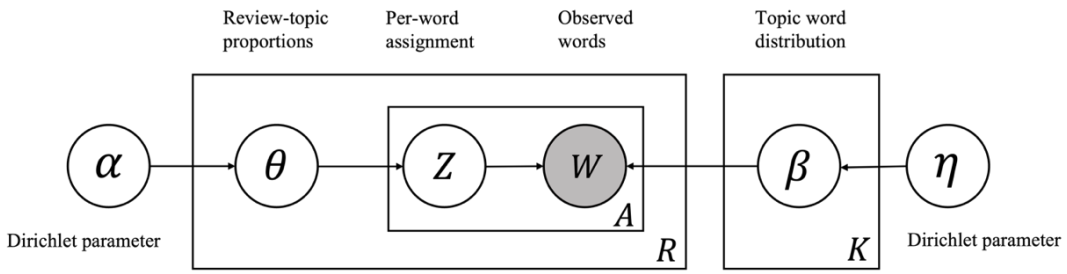

Figure 7 LDA model process using plate notation (Adopted by Blei et al., 2003)

The topic-word distributions and the coefficients for documents cannot be observed and are estimated by a learning algorithm, such as expectation propagation (a higher-order variational algorithm) and the Markov chain Monte Carlo algorithm (Griffiths and Steyvers, 2002; Minka and Lafferty, 2002; Porteous et al., 2008). For model inference and parameter estimation, we adopted Gibb's sampling to compute the approximations to the posterior distribution of the hidden variables in the model, which is the core inferential problem in LDA. Compared with the convexity-based variational approach introduced by Blei et al. (2003), Gibb's sampling could achieve higher accuracy by approaching the asymptotically correct distribution (Porteous et al., 2008).

3.3.4 Supervised machine learning techniques in the classification

To examine the ability of two approaches in predicting rating scores, we conduct a set of two-stage experiments: binary classification and rating prediction. Classification is used to identify which category the new observation belongs to, while prediction involves making future estimations based on current data behaviour patterns (Brintrup, 2021).

SVM

The support vector machine (SVM) is a supervised learning algorithm which is mostly used for classification. It could also be applied to regression problems. The fundamental concept of Support Vector Machines (SVM) is rooted in the calculation of the margin, which is a linear classifier that operates without relying on probabilistic methods (Karmiani et al., 2019; Wu, Huang and Meng, 2008). The margins can be

described as the spatial separation between two supporting vectors that are divided by a hyperplane. The approach makes the assumption that the data exhibits linear separability, allowing for the straightforward determination of the weight associated with support vectors and the optimisation of the margin. In the process of SVM, each individual data element is designated as a point within an n-dimensional space, where n represents the quantity of features. Each feature corresponds to the value of a specific coordinate inside this space. The present methodology is employed for the purpose of analysing vectorized data and identifying a hyperplane that effectively distinguishes between the two input categories. A hyperplane is constructed in a manner that minimises the mean-squared error while maximising the distance between the margin and the classes when the various margins are drawn between classes. Support Vector Machines (SVMs) have the capability to be employed for both classification and regression tasks.

Naïve bayes

The Naïve Bayes classifier, as a type of probabilistic classifiers that utilises Bayes' Theorem of probability in the classification task. These classifiers make significant assumptions about the independence of features. This approach is especially well-suited for situations where the number of dimensions in the input data is considerable. The algorithm relies on the principles of conditional probability. It involves the utilisation of a probability table as the underlying model, which is subsequently updated through the utilisation of training data (Xu, 2018; Yang, 2018; Dai et al., n.d.). The probability table is constructed using the feature values, and it is used to determine the class probabilities for making predictions on fresh observations.

In the context of real-world scenarios, it is unlikely that the assumption of independence among all input features can be satisfied. Naïve Bayes classifiers exhibit distinct operational characteristics due to their adherence to a set of fundamental assumptions, hence justifying their designation as "naïve". The Naïve Bayes model assumes that the predictors are conditionally independent, meaning that they are not related to any of the other features in the model. Additionally, the assumption is made that each attribute contributes to the conclusion with equal weight. Although real-world scenarios often deviate from these assumptions (such as the

dependence of a subsequent word in an email on the preceding word), they serve the purpose of simplifying a classification problem and making it more feasible to compute.

Decision trees

The Decision Tree algorithm is a form of tree-based supervised Machine Learning that is utilised for the purpose of resolving both classification and regression problems. It achieves this by iteratively dividing the data according to a specific parameter. Based on the information contained in the leaves, the judgements are made while the data is divided among the nodes. In the context of classification trees, the decision variable is characterised as categorical, with outcomes typically expressed as binary choices such as "Yes" or "No." Conversely, in Regression Trees, the decision variable is considered continuous, which means it can take on a range of numerical values.

The Decision Tree approach aims to approximate a discrete valued target function (Kotsiantis, 2013; Costa and Pedreira, 2022). The final learned function is represented in the form of a decision tree. The classification process of a decision tree involves the arrangement of instances in a hierarchical structure, starting from the root node and progressing towards specific leaf nodes, based on the values of their features. For each node in the decision tree, it demonstrates a decision or test condition on an attribute of the instance. Meanwhile, each branch in the tree represents a possible value for that attribute. The process of classifying an instance commences at the root node, which is referred to as the decision node. The tree descends along the edge that corresponds to the output value of the feature test, based on the value of the node. The aforementioned procedure persists within the sub-tree governed by the recently added node located at the terminus of the preceding edge. In conclusion, the leaf node represents the classification categories or the ultimate decision.

KNN

The non-parametric technique is commonly employed for both classification and regression tasks. The K-Nearest Neighbours (KNN) approach is employed to

determine the class of an unknown feature vector by identifying the k-nearest neighbours from a set of N training vectors. The algorithm's objective is to categorise a given sample data point into classes when dealing with classification problems. The K-nearest neighbours (KNN) algorithm is considered non-parametric as it does not make any assumptions about the underlying data distribution. It is a robust supervised learning method that operates in a non-parametric and instance-based manner (Zhang et al., 2017; Soucy and Mineau, 2001). One benefit of employing the k-nearest neighbours (kNN) classifier is its ability to consider the localised characteristics of the dataset.

The selection process involves identifying the k-closest neighbours for each entity in the target dataset. Subsequently, the distance between the entity and its neighbours is measured primarily using the Euclidean distance. This assessment helps determine the impact of these neighbours on the classification of the entity under consideration. It is predominantly employed in the field of classification. The K-nearest neighbours (KNN) approach is classified as a lazy learning technique because to its lack of explicit training phase and its reliance on the entire training dataset during the prediction phase. In contrast, it merely retains the data throughout the training phase without engaging in any computational operations. The construction of a model is deferred until a query is executed on the dataset. KNN is considered to be well-suited for data mining purposes.

Random forest

The random forest algorithm is a straightforward supervised machine learning technique that consistently yields superior outcomes without requiring parameter adjustment. As shown in the name, this approach will generate forest and make it more random, which means increasing the quantity of trees within the forest could enhance the accuracy in some way. It can be used to address classification or regression tasks. The Random Forest Algorithm (Breiman, 2001) consists of two distinct steps. The first stage involves the creation of the random forest, while the second stage entails making predictions using the random forest classifier generated in the initial stage. Ensemble learning is a technique employed in both classification and regression tasks. The bagging strategy is employed to generate a collection of

decision trees using random subsets of data. The final choice in a random forest is determined by aggregating the outputs of all decision trees.

The Random Forest algorithm is based on tree-based ensemble structure, wherein each individual tree relies on a set of random variables. The trees utilised in Random Forests are derived from binary recursive partitioning trees. The trees employ a series of binary partitions, sometimes known as "splits," on individual variables to divide the predictor space. The root node of the tree encompasses the entirety of the predictor space. The nodes that remain unsplit are referred to as "terminal nodes" and constitute the ultimate division of the predictor space. Every nonterminal node bifurcates into two descendant nodes, with one node positioned on the left and the other on the right, based on the value of a predictor variable.

XGBoost

Extreme Gradient Boosting (XGBoost) is a highly effective scalable tree-boosting system. It has achieved state-of-the-art results on a wide range of machine learning challenges because of its effectiveness, flexibility, and portability (Chen and Guestrin, 2016).

The ability of XGBoost was compared with a 4-hidden-layers Deep Neural Network (DNN) when making prediction of the team performance (Giannakas et al., 2021). The results revealed that both the learning accuracy and prediction accuracy of XGBoost are higher than DNN. Wu, Li and Ma (2021)applied five different datasets to examine the performances of XGBoost and Multiple-layer Perceptron Neural Network for binary classification tasks. The results demonstrated that XGBoost performed generally better than the neural network and significantly better when the overlapped samples increased. The performance of 7 algorithms were evaluated including Logistic Regression, XGBoost, KNN, Naïve Bayes, Decision, SVM and Random Forests when performing classification task for fake news detection (Khanam et al., 2021). It is examined that XGBoost depicted the highest accuracy than other algorithms.

Table 5 The comparisons of both advantages and disadvantage for 6 different algorithms

Approaches	Advantages	Disadvantages
SVM	• It can handle complex function if the appropriate kernel function can be derived. • Less probability of overfitting because of the generalization adopted in SVM • It can scale up and more effective with high dimensional data.	• Not very good performance goes when dealing with large data. • It is sensitive to noise,which does not work well when dataset is noisy. • It doesn't provide probability estimates as the points are classified above and below the classifying hyperplane.
Naïve Bayes	• Easy implement and training • Capability of good performance with limited training data. • It can also handle both continuous and discrete data • It can be applied to binary or multi-class classification problems. • The ability to generate probabilistic predictions • Insensitivity towards irrelevant features	• Low performance if the independent assumption is not met. • Smoothing required when the probability of a feature turns out to be zero in a class. • Vanishing value due to product of small probability
Decision tree	• Interpretability: Easy in interpretation. • It can easily of handles categorical and quantitative values • It can also fill missing values with the most probable value • Non-Parametric	• Overfitting because it's a high variance algorithm • Optimization: it could be difficult to control size of tree

	• high performance due to efficiency of tree traversal algorithm.	
KNN	• Flexible classification and suited for multi-classes problem • Non-Parametric: ideal for non-linear data since no assumption about underlying data	• The number of K need to be determined • High memory cost and low speed with large dataset • High sensitivity to irrelevant features
Random forest	• Works well for handling outliers. • Works well with non-linear data. • Lower risk of overfitting.	• It could be computationally intensive for large datasets. • Complexity and lower interpretability.
XGBoost	• An in-built capability to handle missing values • It provides various intuitive features, such as parallelisation, distributed computing, cache optimisation	• Sensitivity to outliers • Overfiting is likely to occur in XGBoost

3.3.5 Model performance measurement metrics

In our study, we use two types of metrics to compare the performance of the classification and prediction models. For the classification, a standard approach utilising a confusion matrix is used to represent the dispositions of the test dataset in a 2 x 2 setting (true positive, true negative, false negative, false positive).

A confusion matrix is a tabular representation that arranges predictions based on how they match the actual values. It is a useful instrument for assessing the performance of classification models. As shown in Table 6, matrix representing the prospective categories of the actual value and the predicted values from the classification model.

Table 6 The confusion matrix in the classification task

		PREDICTED LABEL: POSITIVE	PREDICTED LABEL: NEGATIVE
TRUE LABEL	POSITIVE	TP	FN
	NEGATIVE	FP	TN

True Positive (TP) represents the number of positive reviews that are correctly classified into the positive group while Ture Negative (TN) indicates the count of negative reviews which are categorized as the negative reviews correctly. False Positive (FP) represents the number of reviews that originally belong to the negative class but classify into the positive class. False Negative (FN) indicates the number of reviews which positive but classify into the negative class.

Based on the confusion matrix, two performance metrics could be calculated: *precision* and *recall*. They are widely employed in the field of information retrieval to assess the relevance and significance of model results as well as examine the dilution caused by insignificant noise.

Precision is also known as the positive predictive, which quantifies the proportion of correctly identified positive instances, shown in the below formula, which evaluates the accuracy when predicting the positive class. A precise classifier with high level of reliability will only designate the positive class to instances with high probability to be positive.

Precision is calculated as:

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

Recall is a metric to evaluate the completeness of the classification and measure by the number of TF over the total number of positives, as shown in the below formula. The true positive rate, also known as recall or sensitivity, is calculated as:

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

We also adopt F1 score (a singular numeric value combine precision and recall) to evaluate model performance. An equal weighted combination of these two metrics can be reflected on the F1-score, which can be calculated as:

$$F1\ score\ =\ 2 * \frac{Precision * Recall}{Precision + Recall} \tag{10}$$

In addition, another widely used method to examine the trade-off between detecting TF and minimising FP is the Receiver Operating Characteristic (ROC). The ROC curve is a two-dimensional plot that shows sensitivity on the y-axis and $1 - Specificity$ (FPR) on the x-axis.

The false negative rate, also known as specificity, is estimated as:

$$Specificity\ =\ \frac{TN}{TN + FP} \tag{11}$$

The false positive rate (FPR) is equivalent to $1 - Specificity$. The perfect one is located at point (0,1). The ROC curve starts from point (0,0) and ends at point (1,1). AUC is a method to compare classifiers and is calculated as the area under the curve, which is between 0 and 1. The model with higher AUC performs better in classification than others, as it is known that *"the AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance"* (Fawcett, 2006, p. 868).

In addition to the binary classification, we would perform regression using the best algorithm. Thus, to determine the effectiveness of our model, two metrics will be calculated: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). MAE represents the average of the difference between the predicted value and the original value, which can be calculated by the formula below. A smaller MAE indicates a better model.

$$Mean\ Absolute\ Error = \frac{1}{N}\sum_{i=1}^{N} |y_i - \widehat{y_i}| \tag{12}$$

where y_i represents the actual value for the i^{th} instance and $\widehat{y_i}$ is the predicted value by the prediction model.

RMSE is popularly used in the literature, which represents the square root of mean squared error, between the actual and predicted rating score as follows:

$$Root\ Mean\ Squared\ Error = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y_i})^2} \quad (13)$$

All metrics are estimated and used in evaluating the performance of each model in the analysis.

3.3.6 Feature selection

Unlike simple models (e.g., linear regression), more complex predictive models (e.g., deep learning and tree-based models) are complicated to interpret. Shapley Additive explanations (SHAP) values, proposed by (Lundberg and Lee, 2017), could better interpret black-box models by computing Shapley values from coalitional game theory. The Shapley value is an explanation method based on solid theory, in which four axioms (efficiency, symmetry, dummy, and additivity) provide a reasonable foundation. It represents how much a feature contributes to the prediction for each instance compared with the average prediction of the trained model (Molnar, 2020). Inspired by cooperative game theory, the Shapley value is a method of fairly distributing pay-outs to players according to their contribution. In this study's circumstances, the 'players' represent the feature values, and the 'game' signifies the prediction task.

SHAP could explain the Shapley values as a linear model specified as:

$$g(z) = \emptyset_0 + \sum_{j=1}^{M} \emptyset_j z_j \quad (14)$$

where M is the maximum number of simplified input features. $z \in \{0,1\}^M$. When calculating the Shapley value, the value of z represents the status of the presence (used or not used) of the corresponding feature in prediction. $\emptyset_j$ is the Shapley value (the attribution of feature j). The Shapley value of feature j can be calculated as follows:

$$\emptyset(j) = \sum_{s \subseteq \{1,\ldots,M\} \setminus \{j\}} \frac{|S|!\,(M - |S| - 1)!}{M!} (v(S \cup \{j\}) - v(S)) \quad (15)$$

where S represents one subset of the simplified features included in the model. $v(S)$ is the total value for S. The marginal contribution of feature j is calculated as $v(S \cup \{j\}) - v(S)$.

SHAP could be a powerful method to interpret results from tree-based machine learning models (e.g., random forest and gradient boosted trees). In this study, SHAP demonstrates the feature importance by examining its marginal contribution to the model output, which provides local explanation and consistency globally.

3.4 Results

3.4.1 Baseline sentiment calculation

The sentiment polarity score (both document level and sentence level) is calculated using the original textual comment for each review and provide a decimal between -1 and 1. Figure 8 provides the distribution of document-level polarity (left-hand side) scores and sentence-level polarity (right-hand side) in our sample. The biggest peak of the curve is in the middle for two distributions. And the average polarity scores are 0.047 ($sd = 0.228$) and 0.027 ($sd = 0.17$) at document level and sentence level respectively. The most frequently occurring polarity values are clustered near the middle. The extreme polarity scores (close to 1.0 and -1.0) occur the least frequently, which is quite different from the distribution of customer rating scores (Figure 6), as extreme rating scores show the most frequent occurrence.

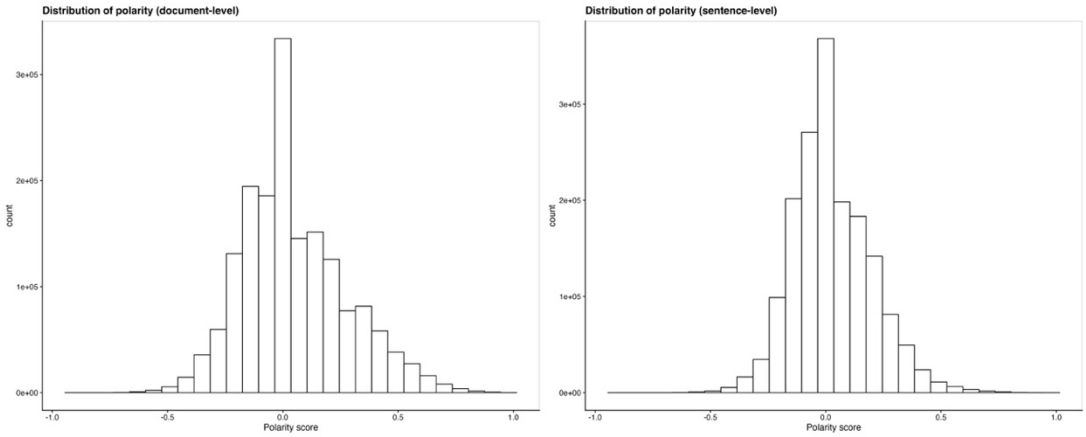

Figure 8 Polarity score distribution for our sample

3.4.2 Corpus pre-processing for topic modelling

The textual content from customer reviews is pre-processed by following the standard procedures, including (a) word tokenization (breaking sentences into a set of tokens), (b) exclusion of numbers and punctuations, (c) elimination of stop words, which

includes both language stop words removal using the SMART stop word list and context-specific stop words exclusion, such as food vocabulary and restaurant brand names, and (d) selecting only adjectives, nouns, and adverbs from remaining words, since these words contain relevant information about products and product quality (Tirunillai and Tellis, 2014). This step is implemented by utilising part-of-speech (POS) tagging to keep the parts of speech that are meaningful as well as lemmatization, which derives the base forms of the words. (e) For low-frequency words (frequency of occurrence is less than 2% of the total number of reviews), a pruning procedure is followed, which reduces the number of reviews to *1,700,131*. The low-frequency words which convey highly specific semantic information are considered as the weak features in the corpus (Leeman, 2007). We followed the procedure of removing low-frequency words based on the previous study (Griffiths and Steyvers, 2004).

3.4.3 LDA model estimation and hyperparameter tuning

After transforming textual data into a document-term matrix, we estimate the topic models and use a heuristic approach to evaluate the hyperparameter values that provide an ideal solution using the current parameter set. As shown in Figure 7, there are two hyperparameters, α and η, which are two smoothing parameters controlling the sparseness of Dirichlet distribution. These two values and the number of topics K are required to be inputted for the LDA process. To determine the best number of K, many researchers have adopted trial and error procedures. A set of models was estimated with various values of K and the model producing the most meaningful topics is selected (Blei, 2012; Bastani, Namavari and Shaffer, 2019). Based on the intrinsic nature of reviews from the OFD platform, we could infer that reviews are homogeneous and concentrated on only a few themes (e.g., food quality, delivery speed, driver's attitude).

The range of topics are determined following two steps. First, we conducted a literature review which employ similar topic modelling approaches to online customer reviews in different hospitality industries including hotels, food takeaway and airlines. Studies which focus on hotels and airlines (Hu et al., 2019; Korfiatis et al., 2019) suggest 20-30 would be an appropriate range for topics. However, because of the intrinsic nature of online food delivery service, the service process is shorter and simplified. In another study (Xu, 2021) which cover online takeaway reviews,7 latent

factors are extracted representing the aspects in a general way. Therefore, to get more specific dimensions and keep the results meaningful simultaneously, a rough range (7-20) is determined in this step. Second, considering the rating mechanism provided by JustEat, ratings are given from 1-6 for 3 different subcategories. The rating score indicates the various levels of customer satisfaction based on their experiences in 3 different subcategories. Therefore, we decided the number of levels (6) of customer satisfaction as the minimum topic number while the maximum value is considered as the number of customer satisfaction multiplied by the number of subcategories, which equals 18. Finally, we adjusted our range of topics to 6-18.

Several researchers seek to find the best number by calculating the *'perplexity'* of the held-out test set, which is one intrinsic evaluation metric for language model evaluation. Perplexity algebraically equals the inverse of the geometric mean per-word likelihood (Blei et al., 2003), which means that the lower perplexity indicates that the model predicts better for new test samples. However, there is a distinct drawback of using perplexity to evaluate the quality of the LDA model. The perplexity decreases as the number of topics increases (Koltcov, Koltsova and Nikolenko, 2014). Chang et al. (2009) illustrated that producing ever finer partitions as the number of topics grew could make the model less helpful and reduce topic interpretability.

Therefore, to find the best number of K, two metrics are calculated, proposed from two studies (Cao et al., 2009; Deveaud, SanJuan and Bellot, 2014) to compare 13 LDA models. Cao et al. (2009) found that the best K is not only correlated with the size of the dataset but is also influenced by the inherent correlations within the corpus. They considered each topic as a semantic cluster, in which the similarity of each word is as small as possible, while the similarities among topics are expected to be large. Similar with the idea of clustering based on density, they aim to achieve a large similarity within the topic for more explicit semantic meaning while a small similarity among topics showing a stable topic structure. The procedures are as follows: First, the initial LDA model is estimated given an arbitrary K value. Second, they calculate the average cosine distance of the model, the model's cardinality, and all topics' density. Third, based on the cardinality, they re-estimate the LDA model and initialise sufficient statistics. If the direction of convergence is negative, topics with high densities will be applied as reference samples. Otherwise, the seeded

method will be adopted to initialise it. Then, repeat the second and third steps until the model's average cosine distance and cardinality converge.

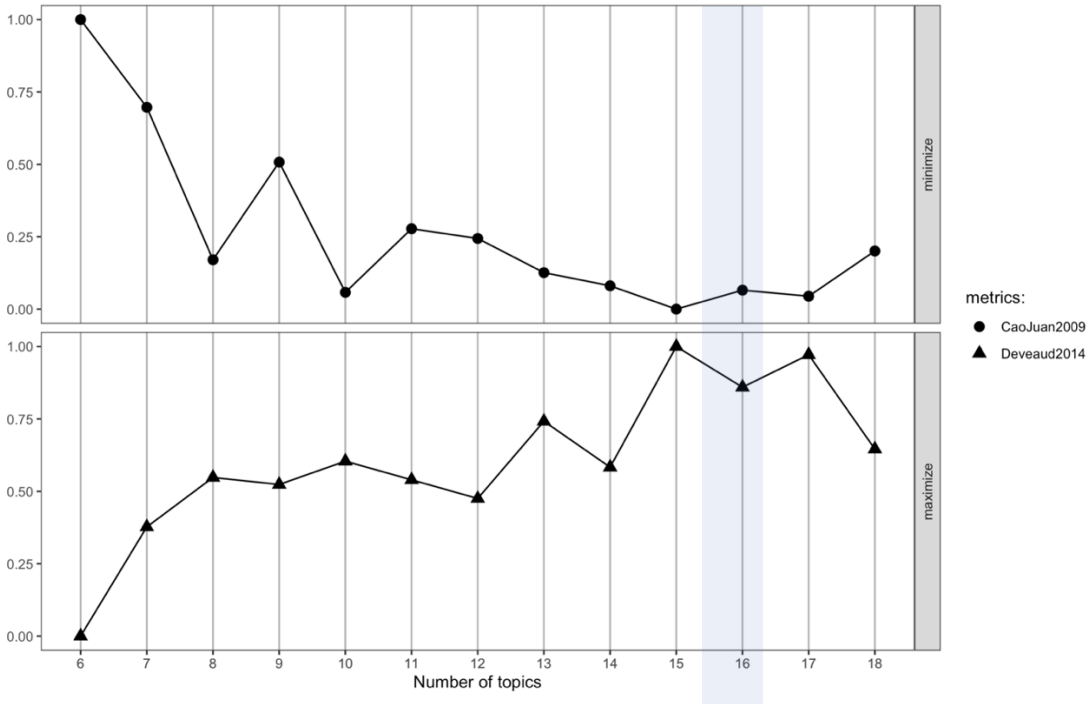

Figure 9 Selection of the number of topics (K) for identifying the topic solution. Optimal K is identified by the shaded area.

In addition, Deveaud, SanJuan and Bellot (2014) proposed a simple heuristic approach to find the best number of topics when the information diverges between all pairs within the LDA model. Rather than the non-symmetric measure (Kullback-Leibler divergence), the symmetrised version of Jensen-Shannon divergence is applied. Figure 9 depicts the performance of different LDA models using these types of metrics for the values of K (x-axis). The model achieves the best performance when the upper metric has the minimum value or the lower one is maximised. Therefore, we select 15 as the optimal value of K.

3.4.4 Topic identification

Table 7Figure 6 provides the K=15 topic solution for the review corpus, which is the optimised solution after topic number selection, as previously discussed. The loading words associated with each topic can be used to understand the main concerns and preferences of customers in relation to the food delivery service being reviewed. The topic solution covers 15 topics' top 7 loading words separately produced using the

standard topic word probability (β) from the LDA estimation process. Several topics show positive intention, such as Topics #1, #2, #3, and #5, while some topics are more negative (i.e., Topics #11, #12, and #13). For instance, Topic #12 mainly talks about the delivery service from drivers about locating their address. Topic #13 concentrates on late deliveries and long waiting times. Topic #3 focuses on customers' praise and subjectively positive descriptions of their takeaways.

Table 7 The 15 topics and their top 7 loading words in the topic solution.

Topic #	**Top 7 loading words**
Topic 1	food, cold, stone, longer, late, delivery, driver
Topic 2	hot, nice, food, lovely, fresh, tasty, again
Topic 3	best, place, amazing, delicious, guy, takeaway, excellent
Topic 4	service, back, customer, phone, poor, bad, problem
Topic 5	food, great, always, early, delivery, fast, home
Topic 6	good, quality, portion, large, small, price, worth
Topic 7	meal, only, happy, extra, box, thing, instead
Topic 8	order, wrong, right, issue, correct, number, store
Topic 9	not, more, disappointed, taste, lot, flavour, same
Topic 10	food, again, warm, free, once, hungry, barely
Topic 11	never, again, money, ever, dry, soggy, hard
Topic 12	delivery, driver, where, door, address, house, man
Topic 13	late, hour, minute, min, half, way, later
Topic 14	time, first, last, long, few, second, next
Topic15	drink, item, missing, bag, refund, order, full

3.4.5 Classification using 6 algorithms

Using Document-level polarity

Table 8 Performance comparisons for all algorithms applied using document-level polarity in the classification task

Algorithm	Precision	Recall	F1-score	F1-score Ranking	AUC	AUC Ranking
NB	0.724	0.850	0.782	4	0.822	4
RF	0.713	0.878	0.787	1	0.812	5
DT	0.724	0.856	0.784	2	0.826	2
XGBoost	0.723	0.855	0.784	2	0.827	1
SVM	0.734	0.824	0.777	5	0.826	2
KNN	0.766	0.437	0.557	6	0.627	6

The precision rate, recall rate, F1 score and AUC score for each algorithm are summarised in Table 8 as well as the rankings for both F1 score and AUC score to better compare each algorithm's performance. From Table 8, it can be observed that the highest F1 score (0.787) occurs for RF with the highest recall rate among all algorithms, followed by XGBoost and DT with the same F1 score (0.784). However, RF performs badly when it comes to the AUC score, XGBoost owns the highest AUC score (0.827), followed by SVM and DT with the same AUC score (0.826). According to the rankings of F1 score and AUC score, it indicates that DT and XGBoost have better performance in classifying positive and negative classes using a single feature (document-level polarity) and achieve relatively better performance than other algorithms. One interesting point is that KNN performs significantly worse than other algorithms. KNN has the lowest F1 score (0.557) because of its lowest recall score (0.437) even with the highest precision (0.766).

Using Sentence-level polarity

Table 9 Performance comparisons for all algorithms applied using sentence-level polarity in the classification task

Algorithm	Precision	Recall	F1-score	F1-score Ranking	AUC	AUC Ranking
NB	0.803	0.909	0.853	1	0.835	5
RF	0.741	0.863	0.798	3	0.837	2
DT	0.741	0.864	0.798	3	0.837	2
XGBoost	0.738	0.870	0.799	2	0.838	1
SVM	0.740	0.849	0.791	5	0.836	4
KNN	0.766	0.598	0.707	6	0.714	6

Similarly, the same algorithms are employed to examine the power of sentence-level polarity in classifying the positive and negative reviews. It can be observed that NB has the highest F1-score (0.853) generated from the highest precision rate and the highest recall rate, representing that NB performs well in identifying positive instance correctly and capturing a high proportion of positive instances among all positive ones. XGBoost achieves the second highest F1-score (0.799) and the highest AUC score (0.838), which indicates its great performance in the classification task using sentence-level polarity. Similar with classification using document-level polarity, KNN has the lowest F1-score (0.707) and the lowest AUC score (0.714) even it achieves higher performance using sentence-level polarity.

Based on the two tables, we found an interesting result that KNN perform much worse than other algorithms using both document-level polarity and sentence-level polarity. It could be affected by the significantly overlapped points in the dataset. As shown in Figure 8, the polarity score with the value of 0 is the dominant group in our sample at both document level and sentence level. It increases difficulty for KNN to classify since 0 is the boundary to separate positive and negative classes. KNN replies on the majority class among the nearest neighbors to assign a label to a new

data point. The existence of overlapped data around the decision boundary leads to the uncertainty in classification.

Generally, both document-level polarity and sentence-level polarity could have incredible ability to classifying the positive and negative reviews. No matter which algorithm is employed, sentence-level polarity plays a more important role in the classification task. Compared with classification using document-level polarity, the sentence-level polarity could promote the F1-score could be from 0.011 to 0.15 and enhance the AUC score from 0.011 to 0.087.

Using Topic membership

The topic membership extracted from the STM result is also utilised in the classification task. We adopt the same 6 algorithms to examine the ability of how topic membership classifies the positive and negative classes correctly. As shown in Table 10, SVM has the highest F1- score (0.789) and the second highest AUC score (0.855). XGBoost has the second highest F1-score (0.785) but the highest AUC score (0.860) while NB performs the worst for both F1-score (0.772) and AUC score (0.840). Also, it could be observed that the performance of KNN is at the same level with other algorithms which is quite different with polarity score previously.

Table 10 Performance comparisons for all algorithms applied using topic membership in the classification task

Algorithm	Precision	Recall	F1-score	F1-score Ranking	AUC	AUC Ranking
NB	0.758	0.786	0.772	6	0.840	6
RF	0.733	0.831	0.779	4	0.845	5
DT	0.752	0.815	0.782	3	0.855	2
XGBoost	0.761	0.811	0.785	2	0.860	1
SVM	0.743	0.841	0.789	1	0.855	2
KNN	0.760	0.792	0.776	5	0.852	4

To summarise, among all the algorithms, XGBoost achieve good performance consistently in classifying positive and negative classes using single feature (polarity score) or multiple features (topic membership of 15 topics). Therefore, XGBoost will be employed and discussed in detail in the following section for both classification and regression task.

Rao et al. (2021) also compared the performance of several algorithms as classifiers including XGBoost, Logistic Regression, Random Forest, Decision Tree, Multinomial Naïve Bayes and Bernoulli Naïve Bayes to perform the binary classification task of detecting fake news. They demonstrated that XGBoost could provide excellent mix of prediction and processing speed simultaneously. After fine-tuning hyperparameters, it could achieve the highest accuracy than other methods. It was examined the power of XGBoost to make predictions in health field. They compared XGBoost with multivariate logistic regression model and found that the former performed better in predicting the risk of death with one specific disease Apart (Yan et al., 2022). Due to the better performance of XGBoost compared with other algorithms, we believe that XGBoost is an appropriate approach for classification and prediction tasks.

Therefore, we selected XGBoost as the optimal algorithm to do the rating prediction. The gradient boosting approach is described as follows: Assume a dataset with n examples and m features $D = \{(\boldsymbol{x}_i, y_i)\}(|D| = n, \boldsymbol{x}_i \in R^m, y_i \in R)$, The output is predicted using K additive functions.

$$\widehat{y_i} = \phi(\boldsymbol{x}_i) = \sum_{k=1}^{K} f_k(\boldsymbol{x}_i), f_k \in F \tag{16}$$

where $f_k(\boldsymbol{x}_i)$ is the predicted value of i-th sample in the k-th tree. Each f_k corresponds to an independent tree structure q and leaf weights w. F is the function space consisting of all CARTs (regression or classification trees in XGBoost), and $\widehat{y_i}$ is the predicted value, for instance i.

The set of functions could be learned by minimising the following objective function.

$$L(\phi) = \sum_{i}^{n} l\left(y_i, \widehat{y_i}\right) + \sum_{k=1}^{K} \Omega(f_k) \tag{17}$$

where l is the training loss function, measuring the difference between the predicted value $\hat{y}_i$ and the observed value y_i. The regularisation term Ω could control the model complexity to avoid overfitting.

The tree model could be trained using an additive strategy. Formally, let $\hat{y}_i^{\,t}$ be the prediction of the i-th instance at the t-th iteration.

$$\hat{y}_i^{\,t} = \hat{y}_i^{\,t-1} + f_t(\boldsymbol{x}_i) \tag{18}$$

Therefore, the objective function at step t is changed as follows:

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{\,t-1} + f_t(\boldsymbol{x}_i)\right) + \Omega(f_t) \tag{19}$$

By using the second-order Taylor expansion, we simplify the equation as below taking the t-step optimization as an example:

$$L^{(t)} \simeq \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}^{t-1}\right) + g_i f_t(\boldsymbol{x}_i) + \frac{1}{2} h_i {f_t}^2(\boldsymbol{x}_i) \right] + \Omega(f_t) \tag{20}$$

where g_i represents the first order gradient statistics of the loss function and h_i indicates the second order gradient statistics. They are represented as follows:

$$g_i = \partial_{\hat{y}^{t-1}} l(y_i, \hat{y}^{t-1}) \tag{21}$$

$$h_i = \partial^2{}_{\hat{y}^{t-1}} l(y_i, \hat{y}^{t-1}) \tag{22}$$

In XGBoost, the complexity is defined as:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda ||w||^2 \tag{23}$$

Where γ and λ are regularisation parameters, T represents the number of leaves and w are scores on leaves. By defining $G_j = \sum_{i \in I_j} g_i$ and $H_j = \sum_{i \in I_j} h_i$ and expanding the regularisation term, the objective function is re-formulated as:

$$O^{(t)} = \sum_{j=1}^{T} [G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2] + \gamma T \tag{24}$$

The best w_j^* and the best corresponding value could be computed as

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{25}$$

$$O_j^* = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j + \lambda} + \lambda T \tag{26}$$

Given that enumerating all possible trees is not intractable, the tree is optimised on one level at a time by splitting leaves and producing a gain score:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \tag{27}$$

3.4.6 Incorporating topic membership in sentiment text detection

As shown in the literature review, sentiment and topic membership could be considered as two devices to explain and predict customer satisfaction. We would like to examine how sentiment and topic membership can predict customer satisfaction empirically. As mentioned in Section 3.1, we consider the mid-point of the rating scale (3.5 stars) as the boundary to separate negative and positive reviews. Based on its rating score, each review in our sample is classified into two classes: *positive* and *negative*. Therefore, by classifying customer reviews into positive and negative, the question is transformed to a binary classification task. More specially, we would also examine the different ability of document-level sentiment and sentence-level sentiment in this classification task. Therefore, we constructed three models (Table 11) with all target variables being rating (positive/negative) for the classification task. Model A and Model B include the calculated polarity score (document-level and sentence-level separately) of each review as independent variables to predict the class. Model C adopted the topic membership of 15 topics from the topic solution of the LDA process as the predictors. The in-sample validation split was 80% for training and 20% for testing with additional folds selected for confidence interval estimation. For both classification tasks, the XGBoost algorithm is followed. For hyperparameter selection, we limit the maximum depth of the tree to 2 and the maximum rounds of boosting iterations to 100 to obtain the best outcome.

Table 11 Three models' AUC Values for the classification task

Target Variable		Polarity (Document)	Polarity (Sentence)	Topic membership (15 topics)
Rating (positive/ negative)	**Model A**	0.827		
	Model B		0.838	
	Model C			**0.860**

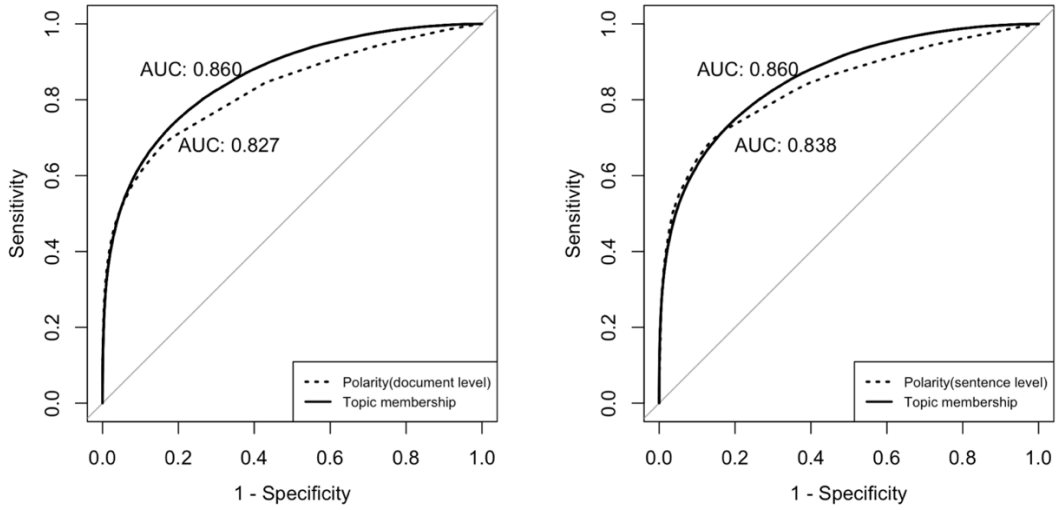

Figure 10 AUC comparison (left-hand side) between Model A and Model C as well as AUC comparison (right-hand side) between Model B and Model C

AUC scores are calculated using k-folding, and ROC curves are graphically presented in Figure 10. As presented Figure 10, the dashed curve represents the ROC curves of Model A and Model B (using document-level polarity and sentence-level polarity separately), whose AUC scores are 0.827 (95% CI: 0.825–0.828) and 0.838 (95% CI: 0.837–0.840), respectively showing relatively good classification, as they are both larger than 0.8. The sentence-level polarity score has better performance than the document-level polarity score in classifying the two rating classes. The solid line is the ROC curve of Model C using topic membership (15 topics), demonstrating better performance than using polarity; the curve close to the upper left corner indicates higher accuracy. The AUC score is 0.860 (95% CI: 0.8587–0.8611), which represents a better ability to separate positive and negative classes. Only comparing the two models with polarity, solely using topic membership on its own could increase the predictive accuracy.

As mentioned, topic membership, including all topics, performs better than sentiment in helping to classify the positive and negative reviews. These 15 latent topics extracted through the LDA process indicate various dimensions customers pay attention to contained in customer reviews. Some of them might be emotional and highly related to customers' ratings, while some might be more realistic. To examine the individual contribution of each topic to the predictive accuracy of binary classification, we construct and perform 15 models (with the same hyperparameters as models in the previous experiment) to classify positive and negative classes and include the topic membership for each topic as predictors. Table 12 demonstrates the AUC scores for the 15 models. Several models (Models #4, #7, #9, #10, #12, and #14) have relatively low AUC scores (close to 0.5) and do not help much in predicting customer attitudes. The remaining models have higher predictive performance for classification in some way, whose AUC scores are higher than 0.6. Among them, Model 3 has the highest AUC score (0.706) and presents the best performance in the binary classification task. The dimension that Topic #3 mainly talks about indicates the strongest relationship with customers' attitudes (positive or negative).

Table 12 AUC Comparison among 15 topics' memberships in the classification task

Predictor	Topic1	Topic2	Topic3	Topic 4	Topic 5	Topic 6	Topic7	Topic8	Topic9	Topic10	Topic11	Topic12	Topic13	Topic14	Topic15
Model 1	0.621														
Model 2		0.677													
Model 3			**0.706**												
Model 4				0.571											
Model 5					0.637										
Model 6						0.612									
Model 7							0.546								
Model 8								0.605							
Model 9									0.548						
Model 10										0.539					
Model 11											0.618				
Model 12												0.542			
Model 13													0.649		
Model 14														0.551	
Model 15															0.626

We already examined the model with topic membership and model with sentiment only separately and proved that topic membership with all topics included could perform better than polarity only in classifying positive and negative classes. However, the combined predictive power of sentiment and topic membership has not yet been examined. As Topic #3 contributes most to the predictive accuracy of classification, we construct Model D and Model E with the integration of Topic #3 and polarity (two levels) as predictors for the classification as shown in Table 13. The former includes document-level polarity score and topic membership (only Topic #3) while the latter includes sentence-level polarity score and topic membership (only Topic #3) as predictors.

Table 13 Two models construction with combination of sentiment and topic (3) membership

Target Variable		**Topic membership (only Topic 3)**	**Polarity (Document level)**	**Polarity (Sentence level)**	**AUC**
Rating (positive/ negative)	**Model D**	●	●		0.842
	Model E	●		●	0.852

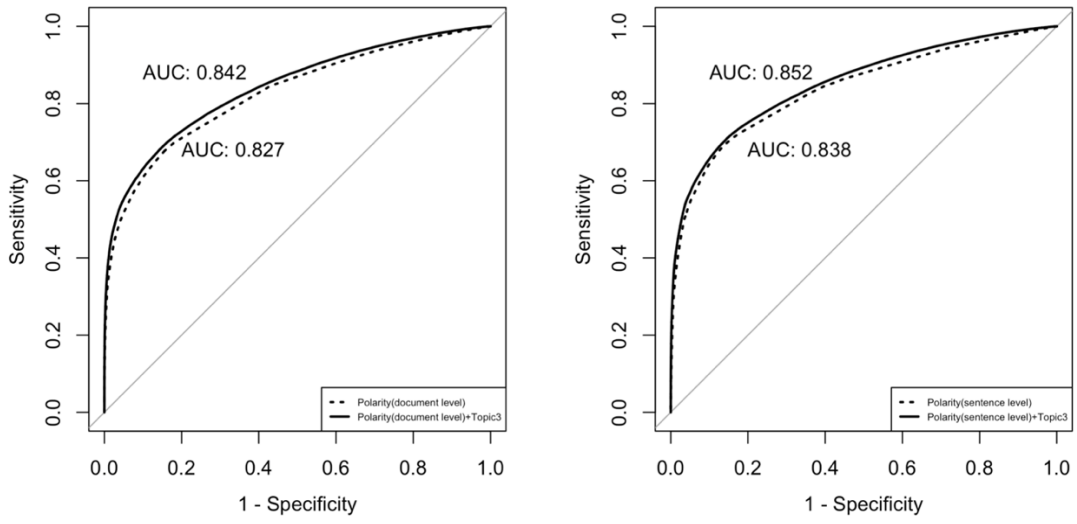

Figure 11 AUC comparison (left-hand side) between Model A and Model D as well as AUC comparison (right-hand side) between Model B and Model E

Models were performed with the same hyperparameters, and their ROC curves and AUC scores are displayed and compared with the Model A and Model B, as shown

in Figure 11. The comparison (the left-hand side) of Model A and Model D showed that Topic #3 added as a new predictor together with document-level polarity could increase the AUC score to 0.842 (95% CI: 0.841–0.843). The comparison (the right-hand side) of Model B and Model E revealed that Topic #3, together with the sentence-level polarity, could increase the AUC score to 0.852 (95% CI: 0.851–0.854). It demonstrates that topic membership possesses the additional power to help sentiment when predicting customer attitudes.

The parameters that decide the model architecture are hyperparameters. We could find the ideal hyperparameters and improve our predictive accuracy through using grid search. In this study, we aim to examine the if the model performance can still be improved or not instead of finding the optimal hyperparameters. Therefore, among all hyperparameters, five hyperparameters are selected for grid search based on the literature and industrial practice including max_depth (the maximum depth of a tree), min_child_weight (the minimum sum of weights of observations required in a child), gamma, subsample (the fraction of observations to be random samples for each tree) and learning rate. We set ranges for each hyperparameter based on the typical values, which generate 240 combinations. After running all combinations, we find the best combination of hyperparameters with max_depth as 7, and min_child_weight as 5, gamma of 0.2, the subsample ratio as 0.8 and a learning rate of 0.3.

Finally, as Figure 12 shows, we improve the AUC scores of models with polarity (both document-level and sentence-level) and Topic #3 as predictors to 0.845 (95% CI: 0.843–0.846) and 0.856 (95% CI: 0.854–0.857), respectively.

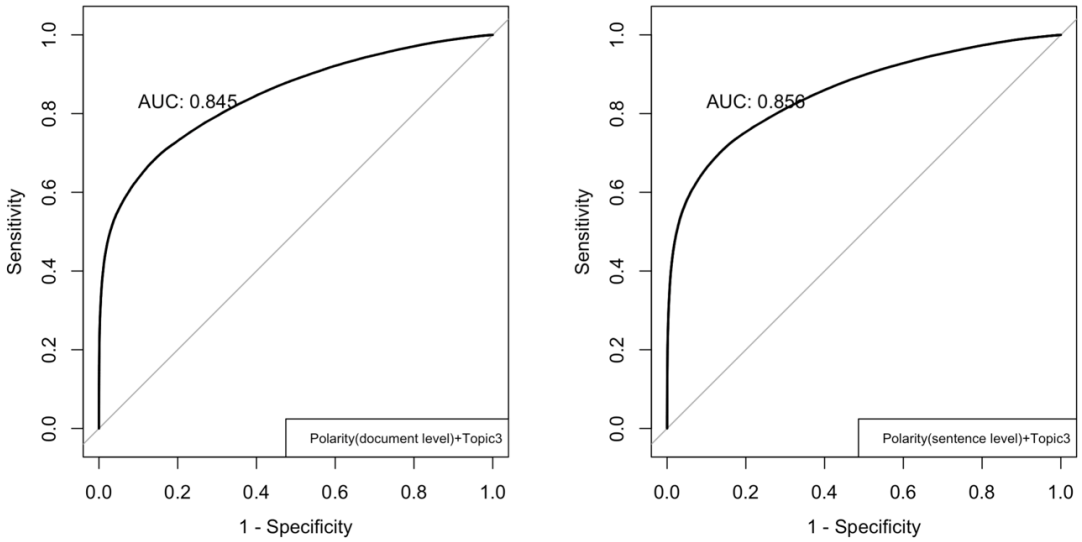

Figure 12 ROC curves for Model D and Model E after tuning the hyperparameters of XGBoost

3.4.7 Rating score prediction

To examine the ability of polarity and topic membership to predict the exact rating score, we construct the baseline model, which includes the (sentence-level) polarity score of each review as the only predictor. The sentence-level polarity will be adopted in the rating score prediction, considering that it is more accurate than document-level performance from the results in the previous section. By combining each topic membership separately with polarity score, 15 models are formed and trained using the same training dataset used in the binary classification task and performed using XGBoost. Considering that this task is an actual prediction task, MAE, and RMSE models are used for evaluation.

We calculate MAEs and RMSEs for 15 models as well as the baseline model. Figure 13 displays the relative difference of MAE and RMSE for 15 models compared with the baseline model and sorts from the highest change to the lowest change. All topic membership could decrease the error compared with the baseline model, which indicates that the inclusion of even one topic as a covariate could increase the accuracy of the prediction task. There are two distinct variables (Topics #13 and #3) that can dramatically decrease the MAE and RMSE.

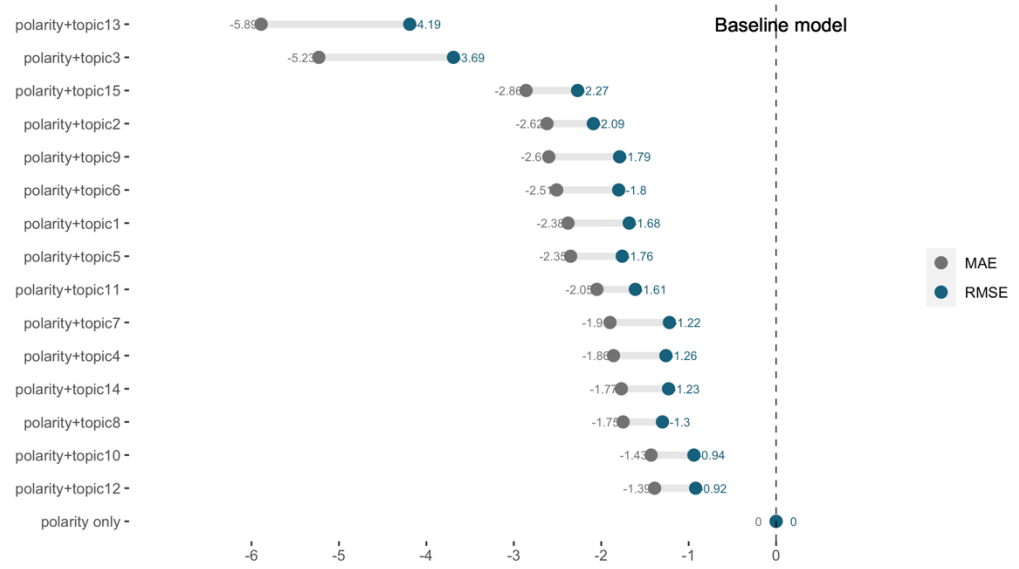

Figure 13 Relative difference of MAE and RMSE for 16 models compared with the baseline model.

3.4.8 Shapley Additive explanations (SHAP) for each feature

While Topic #13 and Topic #3 perform best in reducing the model error, other topics (i.e., Topics #15 and #2) also show improvement compared with the baseline model. To get more specific and direct comprehension of how much each topic can contribute to the prediction of customer ratings, we construct a model including polarity score (sentence-level) and proportions of 15 topics as covariates to predict customer ratings. The accuracy is improved substantially (MAE=0.869, RMSE=1.137) by including all topics. The contribution of each feature to the target value is represented by Shapley additive explanations for feature importance, which is calculated as the average of the absolute Shapley values for each feature across the dataset. SHAP values can be used to understand the relative importance of different features in a model (Lundberg and Lee, 2017). These values are calculated by comparing the model's output with the expected value of the model's output over the entire distribution, taking into account the dependencies between features. By using SHAP values, we can identify the most important features in a model and understand how they contribute to its prediction (Futagami et al., 2021).

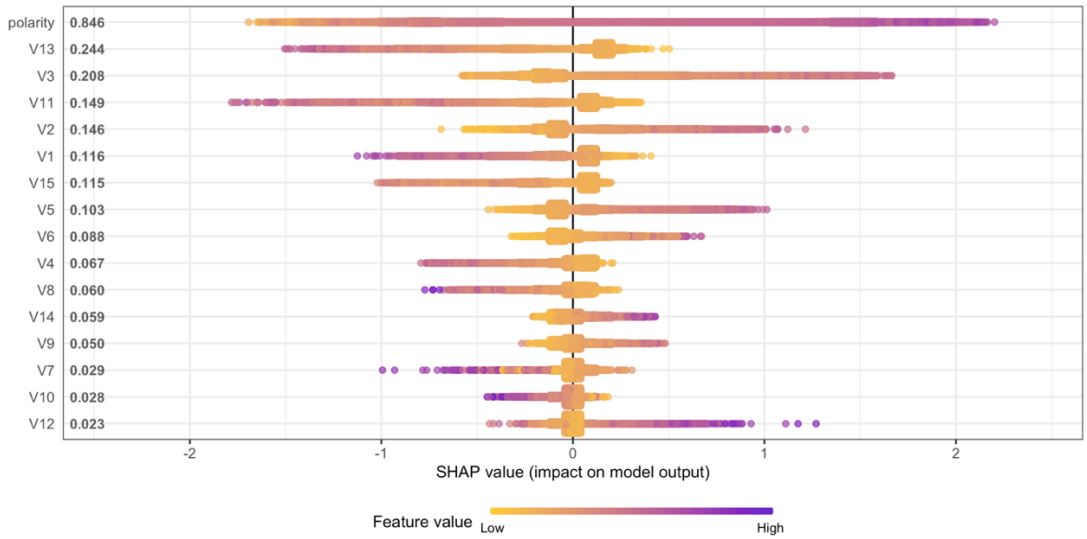

Figure 14 SHAP summary plot

The summary plot Figure 14 displays global feature importance, as well as feature effects. All variables are sorted by decreasing feature importance along the y-axis, with their corresponding value next to them. The polarity score is the most dominant feature, and Topic #13 is the second most important feature, followed by Topic #3, while Topic #12 contributes the least to the predicted values. Each dot in this plot shows the Shapley value of an instance for each feature, whose horizontal location displays its Shapley value, and vertical location is determined by the specific feature. The gradient colour demonstrates the original value for that variable from low to high. Polarity affects the target variable positively, as high polarity scores could increase the predicted customer ratings. Topic #13's membership is negatively associated with the target value as the predicted rating score will decrease while the proportion of Topic #13 within a review increases.

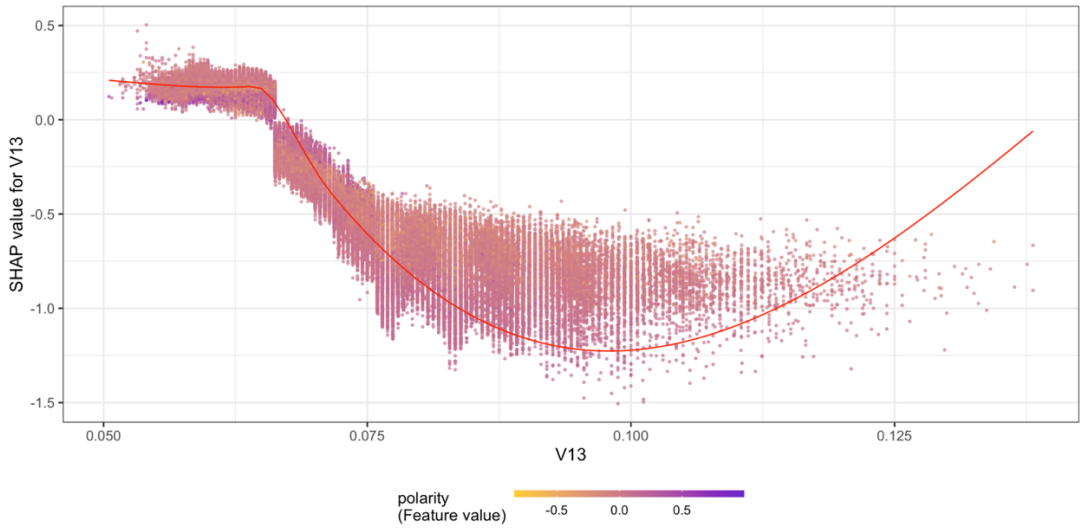

Figure 15 SHAP dependence plot for Topic #13 membership and its interaction visualisation with polarity score

Figure 15 displays the dependence plot for Topic #13 and its interaction with the polarity value. Each dot represents an instance with its proportion within a single review on the x-axis and its corresponding Shapley value on the y-axis. The gradient colour shows its polarity value. A small number of dots with proportions lower than 0.067, have positive SHAP values, indicating an increased prediction value. In contrast, a great many instances have a higher topic proportion between 0.067 and 0.096. Their corresponding SHAP values are lower than 0, meaning that they decrease the predicted value. More explicitly, for these dots whose x-axis is between 0.067 and 0.096, as the proportion of Topic #3 increases, their negative influence on the predicted rating score will be stronger.

3.5 Discussion and implications

3.5.1 Discussion

Through the robust experiment (binary classification), the results show that both document (review)-level polarity score and sentence-level polarity score perform well in classifying a review positive or negative with AUC scores—with both being higher than 0.8 even they are included as the only covariate in two models, respectively. Compared with document-level polarity, sentence-level polarity performs better in the classification task, with a higher AUC score (0.838) than the other AUC score (0.827). Polarity within customer textual content could excellently predict customer

satisfaction, while sentence-level polarity has a better ability for the prediction, which identifies the impact of different granularity levels. It confirms the strong ability of sentiment to explain customer ratings, consistent with previous research (Chatterjee et al., 2021; Zhao, Xu and Wang, 2019).

However, the topic memberships of 15 latent topics extracted from review text using a topic modelling approach (LDA) perform better in the classification task, with an AUC score of 0.860. Topic membership (all topics included) has higher accuracy than the classification task's polarity score (both document-level and sentence level). It represents the multidimensionality exists in customer reviews in which they discuss their opinions towards the food and delivery service from various aspects. The multidimensionality is consistent with other studies which focused on customer reviews from other hospitality industries such as airlines, restaurants, and hotels (Büschken and Allenby, 2016; Xu, 2020). In addition, the multidimensionality not only stands for various entities (e.g., food quality and delivery service), but also demonstrates customers' dialectical in textual content. Customers might describe their experiences dialectically, such as two-sided reviews (Wang, Zhong and Liu, 2022). That may explain the stronger ability of topic membership to predict ratings as both document-level and sentence-level sentiment could not capture the various dimensions (Birjali et al., 2021).

By examining each topic membership to the classification task separately, the one with the highest AUC score is selected and collaborated with the polarity score (two different levels), improving the AUC scores to 0.845 and 0.856, respectively, after hyperparameter tuning. The robust check could prove that the features generated from the textual content could be combined with the sentiment to achieve higher accuracy. It also reveals the heterogeneity of the latent dimensions (Büschken and Allenby, 2016; Hu et al., 2019), which don't equally contribute to customer's overall ratings.

Therefore, we included polarity (sentence-level) and each topic membership one by one as covariates and constructed 15 regression models. Compared with the baseline model (with only sentence-level polarity score), the results indicate that whichever topic membership could improve the model performance. Among them, two topics (Topic #13and #3) could add the most accuracy to polarity in predicting rating scores since the MAE and RMSE of that model are decreased most compared

to the baseline model. The information captured by the two topics could add more predictive accuracy to sentiment for rating prediction, which is also proved by the SHAP feature importance when we include all topics' membership and polarity (sentence-level) into the prediction.

3.5.2 Theoretical implications

Many studies have examined the power of sentiment and latent dimensions within review text to explain and predict customer overall satisfaction (Cheng et al., 2018; Xing et al., 2019; Y. Zhao et al., 2019). Compared with most previous research, our study reveals the comparison the two approaches and how they can be combined for rating prediction.

First, our findings suggest that compared with sentiments, the topic memberships of latent dimensions generated from the review text have better performance in rating prediction. The top membership could be considered as a helpful tool to predict customer overall satisfaction. Furthermore, the heterogeneity of the predictive power exists across different dimensions. Several dimensions have good performance in rating prediction while several dimensions don't show good enough performance. The new features we extracted from review text can be collaborated with customer sentiment to achieve better performance in rating prediction both holistically and individually. It extends the literature by combining the topic membership with customer sentiment instead of only adopting one feature from the review text.

Second, we employed 6 algorithms in the classification task to do the robust check, which increase the validity of our model. It further proved the applicability of our model based on the result of KNN, which result from the highly repetitive numeric value of polarity score, which is the disadvantage of lexicon-based approach. Because of the length and the characteristics of lexicon-based approach, the polarity score derived from the textual data could not achieve high accuracy when the classification algorithm is not adequately selected.

Third, several models have been proposed by researchers to achieve higher accuracy in rating prediction (Cheng et al., 2018; Xing et al., 2019). Compared with these approaches, our approach has more flexibility and interpretability, especially when considering points of intervention in the customer facing areas of the business. The use of topic modelling approach allows us to extract latent dimensions from the

textual data without pre-labelling the data, which saves human efforts of training and adjusting the model. The adoption of unsupervised machine learning approach (LDA) saves the human effort to train the model. In addition, the adoption of Shapley value could clearly show how each dimension extracted from review text contributes to predicting the overall ratings, which provides more interpretability of how each latent dimensions and customer sentiment affect the customer overall satisfaction.

3.5.3 Practical implications

Customer ratings are direct measurements of customer overall satisfaction while the textual content represents customer perception towards experience showing customer satisfaction indirectly (predicting overall). Findings from this study could provide several managerial insights for restaurant owners and managers. First, our findings provide restaurants with identification of latent dimensions from a large amount of customer reviews, which demonstrates customers' various aspects of perception towards food and delivery service. Both praise and complaint from customers could help restaurant owners to develop operation strategies.

Customer feedback is a vital information source to understand customers' opinions, which has been proved to have significant influence on customer behaviour and sales (Li et al., 2019; Z. Zhao et al., 2019). The 15 topics extracted from review text shows the multidimensionality of customers' evaluation. For instance, Topic #13 and Topic #8 illustrates customers' complaint about long delivery time and order issues respectively while Topic #2 represents customers praise towards food quality.

Furthermore, the heterogeneity which exists in the contributions of each topic membership to customer overall satisfaction, may help restaurants to prioritize the most influencing factors. By identifying the most important positive and negative dimensions that influence the rating scores, restaurant managers could explicit dissatisfaction and enhance their weakness accordingly or develop marketing strategies by highlighting their strength. For instance, our findings show that except for polarity, Topic #13 mainly illustrating the long delivery time contributes most to predicting rating score. And the long delivery time is the most important factor negatively affecting the rating score minimizing the reviews lead to higher rating score. Following an identification of the contribution of each service quality factor, an owner may also identify priorities and monitor improvement by incorporating

other covariates such as service time and other inputs such as quality of raw materials etc.

3.6 Conclusions, limitations, and future research

Finding the underlying reasons for rating scores using contextual information is critical for businesses to develop strategies to discover why customers have different levels of satisfaction. To discover the value of unstructured text within customer reviews to explain actual customer ratings, we evaluate and compare how two approaches (*sentiment analysis* and *topic models*) can be applied to understand customer satisfaction. This study demonstrates that incorporating document-level covariates, such as topic membership, can greatly contribute to the understanding of the sentiment of customer feedback, such as the ones found in customer reviews, and predict the review score in a much better way than the actual tone of the review text, even in cases where access to sentiment vocabulary may be limited. While a large body of literature has demonstrated that review text is primarily consistent with the rating—and therefore, the sentiment of the review is reflected in the star rating of this particular review, from a business owner's point of view—latent dimensions discovered from these review texts are a useful instrument to be incorporated in the business practice. They can identify areas of improvement and competency that the business can expand the most.

Nevertheless, there are several limitations to our research. For the robustness check in our first experiment, we only classify reviews into positive and negative classes, not considering the neutral class, which has been commonly studied in online review literature. Furthermore, even though we have examined two granularity levels of sentiment, we only employ a single dictionary. The choice of dictionaries may have a different influence on prediction accuracy. Therefore, these aspects could be improved in future work.

Future work should focus on several directions. First, reviews classified as neutral could be considered, together with the positive and negative ones, as three distinct classes for classifying customer rating scores to different satisfaction levels. Second, apart from lexicon-based methods, other unsupervised machine-learning techniques could be adopted to detect customers' polarity scores. Also, the subjectivity and emotions contained in the review text could be considered in future work. Third, customer reviews from other platforms could be included in the future

to discover the level of heterogeneity across different platforms and different service domains.

CHAPTER 4 Identifying latent service dimensions in service supply chains: an analysis of franchised and independent businesses

4.1 Introduction

In recent years, services have assumed an increasingly important role in the economy across many countries. For example, in the UK, according to the World Bank, the service sector accounted for the majority (approximately 72.8%) of the gross domestic product (GDP) in 2020 (World Bank, 2021). For service businesses, the role of customers is particularly important due to their high involvement in the service delivery process (Sampson and Spring, 2012). In that regard, business owners strive to increase customer satisfaction by improving and maintaining service quality (Hunneman, Verhoef and Sloot, 2015) while addressing the challenge of bundling products with services. The processes involved in bundling physical products with intangible services, from their point of production to their point of consumption, comprise the *service supply chain*.

There are three critical processes (order process management, service performance management, and customer relationship management) that need to be effectively managed to ensure a smooth relationship between the service provider and its customers (Baltacioglu et al., 2007). However, when products/services flow from the service providers to customers, there could be several issues, which could affect the final service quality and customer satisfaction (e.g., wrong specification, late delivery etc.). Therefore, it is necessary to identify these issues and gauge any potential insights for business owners.

Across service supply chains, there is a high degree of heterogeneity in terms of the organisational form of service provision. One of these forms is *franchising*, a strategy adopted by many firms, with primary objectives being market positioning and maintenance of quality standards in product and service delivery (Gillis, Combs and Yin, 2020; Sawant, Hada and Blanchard, 2021). In the UK, the franchise industry's revenue has steadily expanded from only £1,3 billion in 1985 to £15,1 billion in 2015. Personal services, hotels and catering continue to be the fastest-growing franchise sectors (British Franchise Association, 2017;2018).

Franchising helps franchisors to rapidly expand, by drawing from 'local' managerial expertise (Gillis, Combs and Yin, 2020; Norton, 1988), reducing monitoring costs and enhancing profitability (Hsu and Jang, 2009). Franchisees are expected to follow standardised processes and deliver services of standard quality on behalf of the franchisor under a shared trademark. However, due to their "residual claimant" status, there is a risk that franchisees will free ride by reducing the quality of products/services in order to increase their own profit margins. This harms the reputation of other franchise units as well as the entire franchise system (Kidwell, Nygaard and Silkoset, 2007). An owner of a small service firm can seek to grow its business by becoming a franchisee of a large chain. As a franchisee, the firm can get access to the franchisor's supply chain and achieve economies of scale and scope through obtaining goods and services from franchisor-vetted suppliers, potentially enhancing product/service quality (Combs et al., 2011). Besides, standardised and centralised operations enhance efficiencies and might generate competitive advantages over independent rivals in the local market (Sorenson and Sørensen, 2001). Moreover, the franchisee adopts the franchisor's brand name, an intangible asset and one of the most valuable ones for every firm (Keller and Lehmann, 2003).

This means that it can generate revenues due to customers associating it with a recognisable brand. However, the firm can choose to remain independent as independent businesses have the freedom to modify products/services based on marketing conditions or personal preferences and tailor operational and marketing strategies without the restriction from franchisors (Kaufmann, 1999; Kuratko, 2016). Despite the heterogeneous, intangible, and inseparable nature of service outputs (Parasuraman, Zeithaml and Berry, 1985) as well as the multidimensionality of customer satisfaction in service encounters (Tirunillai and Tellis, 2014), it is reasonable to expect that there will be discernible aspects of product/service delivery that 'matter' to the customer. The question is to what extent these aspects differ between the two forms of organisation for service provision – independent versus franchisees – and whether the two perform comparatively better or worse across these dimensions from a customer's point of view. We thus aim to answer the following research questions: (a) *What are the latent dimensions affecting customer satisfaction during the process of producing and delivering products/services to end-customers? (b) What are the differences across these dimensions between franchised and independent businesses?*

We answer these questions in the context of online food delivery (OFD) services. This choice of context is driven by three reasons: First, the franchising model has been prominent and adopted for a long time in the restaurant industry, and increasingly, with the advent of e-commerce, most restaurant chains have entered the OFD sector. With the adoption of OFD platforms, this creates a new competitive landscape. Second, the recent COVID-19 pandemic has increased the demand for food ordered online, so many independent restaurants have attempted to establish a stronger online presence through the use of OFD platforms. Third, OFD platforms are rich sources of unstructured data (numeric ratings, textual reviews) that allow for evaluating aspects of customer satisfaction controlling for restaurant type as well as taking into account other factors such as cuisine and location when sampling the data.

To this end, we collect a large corpus of online customer reviews (N = 553,807) from a total of 10,894 restaurants (franchised and independent) on *JustEat*, the dominant food ordering platform in the UK. Online reviews, a typical form of electronic word of mouth (e-WOM), have been utilised by firms for marketing purposes to inform and persuade consumers (Ruiz-Mafe, Chatzipanagiotou and Curras-Perez, 2018). However, the free text provided by reviewers contains information content that can be used to develop or improve management strategy (Pantano, Dennis and Alamanos, 2021). We utilise these textual comments (*unstructured data*) to identify the latent dimensions of customer satisfaction with services. Our chosen topic modelling approach allows us to model the relationship between individual dimensions and review ratings (*structured data*), and the role of ownership type.

This work contributes to a more comprehensive study of franchising outcomes, by focusing on the customer's perspective, an issue that has been overlooked by the relevant literature. The results reveal the latent dimensions extracted from customer reviews, and their relationship with review ratings. From a practical standpoint, this study provides insight to managers of establishments in the OFD sector, to help them identify the main issues encapsulated in the process of producing and delivering products and services. Managers in charge of large franchising systems, as well as independent entrepreneurs wishing to outperform competitors, can be equally benefitted from this study.

Our study is beneficial for service supply chains that are exploring ways to incorporate customer feedback into their operational improvements. The significant

power of using online customer reviews to enhance the service supply chain in internet services is highlighted by Rajendran and Fennewald (2021). They suggest that valuable insights obtained from customer feedback can have a substantial impact on supply chain strategies in the service sector. Our study align with their study as customer and their opinions are considered as the vital factors in the service production process (Shahin, 2010). In the OFD service supply chain, the latent dimensions extracted from customer opinions could be applied to better service supply chain management as they are connected with the three critical processes that need to be effectively managed to ensure the smooth relationship between the entities within the service supply chain.

In addition, there are service supply chains from other industries that could also benefit from this study, such as online retail, hospitality, and other internet-enable services, which heavily rely on customer online behaviour. This study analysed the customer feedback from the demand-side perspective by extract latent dimensions of customer satisfaction and the diversity of these dimensions across two types of businesses. It can directly inform supply-side decisions, including service improvement initiatives and improving market positioning strategies as well as operational effectiveness (Li et al., 2021).

The rest of the paper is organised as follows: Section 2 presents the theoretical background of this paper, including the service supply chain in the OFD sector, the role of franchising, and the utility of online reviews. Section 3 illustrates our data sample and research methods. Section 4 presents and explains the results. Section 5 concludes our research and provides theoretical and managerial implications. Section 6 states the limitations and proposes future research directions.

4.2 Theoretical background

Our research draws from three research streams. The first concerns the dynamics of the service supply chain in the OFD sector, and the key characteristics that affect consumer attitudes. The second relates to the role of franchising as a mode of service provision. Finally, the utility of online reviews is demonstrated through relevant applications.

4.2.1 Service supply chain management in the OFD sector

With the growing size of the service sector, there is an increasing demand for deeper knowledge and more effective administration of services. As such, academics seek to understand how services can be effectively managed in various types of service supply chains. Prior studies have concentrated on aspects such as outsourcing, logistics, and service capacity (Ghosh, Lee and Ng, 2015; Jain, Hasija and Popescu, 2013; Wei, Hu and Xu, 2013). Additionally, theoretical frameworks have been constructed to highlight the critical activities and processes of service supply chain management (Baltacioglu et al., 2007; Giannakis, 2011) while various models have been built that try to measure and evaluate the performance of the service supply chain (Boon-Itt, Wong and Wong, 2017; Tseng et al., 2018). The service supply chain is defined as a network of suppliers, service providers, and customers, as well as supporting units, through which resources are transformed into services and delivered to customers (Baltacioglu et al., 2007). There are two broad types of service supply chains: SOSCs (Service *Only* Supply Chains) and PSSCs (*Product* Service Supply Chains) (Wang et al., 2015). SOSCs are those in which no physical objects are delivered, and all outputs are intangible services. These exist in a variety of industries such as healthcare, psychology consultation, etc. In PSSCs on the other hand, alongside the intangible services, there are tangible products that comprise key aspects of the offering. PSSCs are prevalent in the hospitality sector that includes restaurants and beverage retailing (cafes, bars).

In the OFD service supply chain, there are several key elements during the whole process of producing and transporting products/services to customers. Based on the dimension of tangibility, they could be summarized into two aspects: tangible aspect and intangible aspect. The restaurants provide a bundle of goods (food) and services (order processing, delivery), giving rise to PSSCs. As to the tangible aspect, food production is the fundamental and crucial activity. Food quality refers to how well the food performs in meeting customers' expectations, and is a critical factor determining the customer experience, as well as a restaurant's operations strategy. Numerous studies have examined the significance of food quality in the restaurant industry. Numerous literatures have examined the importance of food quality in restaurant industry. Customer loyalty could be affected by food quality as food quality is a fundamental factor in the whole food delivery process. For example, Sulek and

Hensley (2004) demonstrated that food quality, rather than service quality or the physical atmosphere, is the primary driver of customer satisfaction. Another study proposed six attributes of food quality, including healthier option, taste, freshness, presentation, temperature and variety, and examined their effect on satisfaction and purchase intention, highlighting taste and presentation as the most important aspects (Namkung and Jang, 2007). Additionally, Konuk (2019) identified a positive influence of food quality on customer perceived value in organic food restaurants specifically.

When it refers to the intangible aspect, service is the main component that makes the service supply chain different from the traditional manufacture supply chain. Intangibility is the key distinguishing feature of services (Parasuraman, Zeithaml and Berry, 1985). It is emphasized that service supply chain played an increasing important role because of the fast-growing scale of services-sector business. It is not possible to touch, smell, taste, or see a service, as it is a performance rather than a tangible object (Ellram, Tate and Billington, 2004).

Services are thought to possess four defining characteristics that distinguish them from products: intangibility, simultaneity, heterogeneity and perishability (Baltacioglu et al., 2007). In the restaurant industry, Ha and Jang (2010) demonstrated that service quality has a positive and significant influence on customer satisfaction and loyalty, and emphasised the moderating effect of atmospherics in restaurants. It was showed that service quality is a substantial predictor of restaurant image, which in turn has a significant impact on consumer perceived value. However, in the OFD sector specifically, service quality concerns almost entirely the order preparation and delivery process. Issues with these (e.g., delays, wrong order delivered) could dissatisfy customers (Ryu, Lee and Kim, 2012).

The relationship was examined between the delivery time and transaction volume and indicted that early deliveries are slightly related to consumer comments while there are no obvious relations between the delivery time fulfilment and the transaction volume of restaurants (Correa et al., 2019). Research showed that a short waiting time is positively related to consumer's level of satisfaction. There are some online food delivery service providers which offer delivery services rather than restaurants themselves while some chain brands prefer to undertake the delivery by their own teams (Polas et al., 2018). Drivers from both platforms and restaurants have a variety of vehicles to choose in order to pick up meals from restaurant and complete the

deliver. Restaurant managers should avoid late delivery and shorten the duration of delays as these lead to customer anxiety (Rao, Griffis and Goldsby, 2011). In addition, the interaction between the person delivering the order and customers is a critical component of service quality. It has been shown that customers' perceived value of the service could be affected by the communication between them and the deliverers (i.e., drivers) as well as the kindness and politeness exhibited by the latter (Payne, Storbacka and Frow, 2008). In summary besides the quality of the food, issues pertaining the processing and delivery of the order are important determinants of customer satisfaction with an OFD business.

4.2.2 The importance of online reviews

As a form of electronic word-of-mouth (e-WOM), online reviews capture consumers' attitudes and feelings about their purchasing experiences. Most online retailers or third-party websites set online boards for consumers to share their views and provide feedback after they purchased or used products/services. This information might enable potential consumers to make more appropriate purchase decisions (Wang and Chaudhry, 2018). Commonly, online reviews contain both numerical ratings and open-ended comments (Park, Lee and Han, 2007). In the comments, users might provide descriptions of the product/service and evaluate their positive and negative features. Their opinions are subjective and could be overall positive or negative (Wei and Lu, 2013).

A substantial body of literature has focused on the characteristics of reviews and reviewers and examined how online reviews affect consumer behaviours and product sales (Chevalier and Mayzlin, 2006; Zhu and Zhang, 2010). In the tourism and hospitality industry specifically, consumers attach great importance to previous customers' reviews. This is because relevant products and services are hard to differentiate, so product and service quality is difficult to be evaluated before the purchase of the offering. This important role of online reviews has been demonstrated in a host of studies. For example, it is demonstrated that the valence of online reviews increases price premium by influencing the seller's credibility and trustworthiness (Pavlou and Dimoka, 2006). Several experimental studies have also shown that the quantity and quality of reviews positively influence purchase decision (Park, Lee and Han, 2007). Online reviews also influence a seller and brand reputation, which have

long-term effect and lead to product/service repurchase (Clemons, Gao and Hitt, 2006; Standifird, 2001).

Online reviews have also been utilised to assist supply chain management processes and activities, such as sales and demand forecasting, and product design and development (Chong et al., 2017; Yang et al., 2019). For example, Chong et al. (2017) showed the predictive ability of the characteristics of online reviews (valence, volume, and helpfulness) for customer demand using data from Amazon. This stream of literature demonstrates the importance and implications of online reviews not only from the perspective of the consumer but also from the perspective of the business that is keen to identify and address issues with their operations and supply chain strategy. We argue then, that businesses operating in the OFD sector can benefit from formulating (or modifying) strategies that respond to aspects of importance to customers.

The prevalence of online reviews has experienced a notable rise in the service sector. The increasing prevalence of social media and online platforms has created a need for the use of online reviews as a means of doing market research and identifying chances for service recovery (Donthu et al., 2021). According to Lo and Yao (2019), the popularity of online reviews is bolstered by the perceived credibility of customer reviews with consistent ratings written by experts. The crucial role of online trust in fostering consumer engagement was demonstrated by discovering how to engage customers using credible and helpful online reviews (Thakur, 2018).

Service providers utilise online reviews as a means to effectively monitor and improve their market presence and performance. The effective management of online reviews, particularly in terms of overall rating and the handling of negative comments, has significant importance in hospitality industry such as hotels as the aspects have been identified as influential indicators of hotel performance (Kim, Lim and Brymer, 2015). Online reviews is employed as a channel to gain insights into customer behaviour and improve the performance of hotels because the factors that contribute to consumer satisfaction could be identified inside hospitality industry (Zhao, Xu and Wang, 2019; Lu and Stepchenkova, 2015). The quantity of online reviews generated by consumers is also examined to have positive influence on restaurant performance.

Online reviews play a crucial role in the service industry, acting as a critical instrument for comprehending customer attitude thus enhancing the service quality. Previous researchers have adopted various approaches to analyse the online reviews

which could provide multifaced insights. The linguistic approach could be employed in order to reveal the veracity of online reviews, which could increase human's ability to differentiate between fictitious and authentic reviews (Wu et al., 2020).

However, a variety of studies adopt various text mining techniques combined with machine learning models to discover online reviews. For instance, Ye et al. (2022) proposed a framework to examine online reviews within the context of the hotel sector. The proposed framework is based on an enhanced k-nearest neighbour model and a latent Dirichlet allocation model. The research examined a dataset comprising more than 8 million customer reviews from a total of 6,409 hotels located in 50 locations across China. The main aim of this study was to utilise online review as a means to improve services, therefore increasing competitiveness in the service industry. The extensive examination of the large dataset of customer reviews yielded valuable insights into the competitive dynamics of the hotel sector. The utilisation of mixed approaches in this study is able to extract significant insights from online reviews, which enables a deeper comprehension of consumer attitudes and preferences. It can contribute to facilitating the formulation of effective strategies to maintain competitiveness within the hotel sector.

4.2.3 Franchising as a strategy

Franchising contributes to the development of the hospitality sector in a significant way and is very common in the hotel and restaurant industries. Capital constraints are common for small businesses, limiting the profitable opportunities that can be followed; franchising can provide such firms with sources of capital and help them become more efficient through process standardisation (Martin and Justis, 1993). Franchising allows firms to expand rapidly by offering them access to several resources (e.g., capital and talent) (Sawant, Hada and Blanchard, 2021). According to agency theory, franchising helps franchisors reduce monitoring costs of units located in geographically dispersed locations (Brickley and Dark, 2003; Norton, 1988).

There are several differences between franchised and independent restaurants. Franchisees might incur higher operational costs, since, besides the initial franchise fee, they need to pay royalty fees to the franchisor (Kuratko, 2016). Additionally, they have relatively less autonomy than independent restaurants, since there are standardised guidelines that need to be followed for restaurant image continuity

(Perryman and Combs, 2012). Nevertheless, by carrying the brand name of the franchisor, they can reduce operational risk and access an established customer base (Chen and Su, 2014; Koh, Lee and Boo, 2009).

Empirical findings on the outcomes of franchising are mixed. Regarding the financial performance, Aliouche and Schlentrich (2009) evaluated the value creation of restaurant franchisors and non-franchisors over a ten-year period (1993-2002). They illustrated that compared with non-franchisors, restaurant franchisors are slightly more likely to generate more economic value and market value. Hsu and Jang (2009) used ROA (Return on Assets), ROE and Tobin's Q value to measure the performance of franchising and non-franchising firms in the restaurant sector over ten years from 1996 to 2005. It is examined that franchising is able to enhance the profitability in the restaurant sector.

There is evidence from the hotel industry that the strategy increases growth and profitability (Enz, Peiró-Signes and Segarra-Oña, 2014; Silva, Gerwe and Becerra, 2017). Similarly, one research found that across five financial performance measures, franchised restaurants perform relatively better than non-franchised ones (Moreno-Perdigón, Guzmán-Pérez and Mesa, 2021). Relevantly, customer satisfaction was examined that customer satisfaction could be increased by chain affiliation (Barthélemy, Graf and Karaburun, 2021). On the other hand, Carvell, Canina and Sturman (2016) found that when compared to brand-affiliated hotel units, independent hotels generate higher revenues per available room.Moreover, Moreno-Perdigón, Guzmán-Pérez and Mesa (2021) demonstrated that they did not identify any significant difference in customer satisfaction between independent and chain-affiliated hotels.

These divergent findings suggest that whether franchised or independent units perform better might depend on the context, suggesting that further research on this topic is justified. Moreover, this question has not been explored much in the context of OFD sector, in which this study takes place. Through textual analysis of reviews, one can identify latent dimensions that determine customer satisfaction, and systematically compare the performance of the two types of businesses.

4.2.4 Research gaps

As a summary of the literature reviews, there are several research gaps that this study is able to fill. First, even the informational value of online reviews has been well discovered in the hotel sector, restaurant sector and online retailing sector, the exploration of customer feedback within OFD sector is still scant. Compared with the restaurant industry in which services are conveyed directly and simultaneously to the customer in the same environment, the services within OFD industry are transmitted through the third party (either brand-own delivery team or drivers from the platform). Therefore, online reviews could act as a useful tool to identify the risks during the service transmission across each node of the service supply chain. Although there are several research that adopted various methods to examine the activities and operations in supply chain of different industries including tourism (Su and Teng, 2018) and healthcare (Ko et al., 2019), the number of studies focused on the OFD sector is still limited. The literature review demonstrates the importance and implications of online reviews not only from the perspective of consumer behaviour but also about identify issues from supply chain activities and developing strategies. It gave us confidence to consider online reviews as a tool to discover the facets in OFD service supply chains in the customer's perspective.

Second, past studies have discovered the outcomes of franchising in several ways, such as the survival rate and financial performance. These divergent findings suggest that whether franchised or independent units perform better might depend on the context, suggesting that further research on this topic is justified. However, the outcomes of franchising from the customer perspective is also vital due to the transition from product-focused strategies to service quality and customer satisfaction has been considered as significant driver of customer loyalty in service industry (Hallencreutz and Parmler, 2021). This study could demonstrate the performance of franchising strategy in from the customer's perspective by the diverse dimensions from customer feedback towards franchised restaurants and independent restaurants. Through textual analysis of reviews, one can identify latent dimensions that determine customer satisfaction, and systematically compare the performance of the two types of businesses.

4.3 Data and methods

4.3.1 Data

Our data is sourced from *JustEat*, the dominant online food delivery provider in the UK. On the platform, customers are encouraged to leave comments to express their opinions and describe their experiences after their orders are delivered. These reviews will display on restaurants' pages which can be referred to by new potential customers who are browsing to select a restaurant.The rating system adopted by *JustEat* is a 6-point scale, with customers invited to provide ratings as well as a textual justification. Each review consists of a rating score and review comment as well as the date when the specific review was published. To make the analysis meaningful, reviews that have only ratings, but not textual comments are filtered out.

We collect customers' reviews from both franchised and independent restaurants from JustEat, which were published between November 2017 and November 2021. The dataset contains the rating score, textual content, review published date, restaurant id, review id, restaurant type and brand name. The dataset contains 553,807 reviews which were casted for a total of 10,894 restaurants.

Additionally due to the text mining process to be miningful, a review length requirement was imposed using a standard winsorizing process of pruning excessively long reviews (higher than 200 words) and very short single phrased ones (15 words). Table 14 provides the characteristics of our sample. There are two types of restaurants: franchised restaurants (45.50%), which are associated with 31 different brands, and independent restaurants (54.50%). Although the average number of reviews per restaurant does not differ between the two types, independent restaurants achieve a statistically significant higher average rating (4.24 versus 3.4; $t(df=4,954) = -48.728, p<0.001)$ and their reviews are longer (35.7 versus 33.64 words; $t(df=4,954) =-12.371, p<0.001)$. The means of the entire sample are 3.86 and 34.77, respectively.

Table 14 Sample characteristics

		Franchised	*Independent*	*Both*
Number of restaurants		4957(45.50%)	5937(54.50%)	10,894
Average number of reviews per restaurant	*M*	52.24	49.66	50.84
	SD	71.49	71.58	71.54
	t(df=10,552)	[1.8713]		
Average rating per restaurant	*M*	3.40	4.24	3.86
	SD	0.84	0.95	1.00
	t(df=10,857)	[-48.728***]		
Average review length (words) per restaurant	*M*	33.64	35.71	34.77
	SD	8.06	9.36	8.86
	t(df=10,882)	[-12.371***]		
Total reviews		553,807		

Notes: Reported t-values for the mean differences between the two groups are based on Welch's unequal variances t-tests.

***: *p-value* <0.001

The rating score distribution for two types of restaurants are shown in Figure 16. After the order is completed, customers are able to rate their experiences from 1 to 6 based on three categories (food quality, delivery time and restaurant service). The final rating score shown on website is calculated as the mean of three sub-ratings, thus creating 16 different levels (1,1.3, … ,6.0).

We used the percentage of each rating score to make it comparable between franchised and independent restaurants. As shown, the extreme rating scores are popular for two types of restaurants. For franchised restaurants, the highest proportion for rating score occurs to 1 star, followed by 2.7 and 6 stars while the percentage of 6 stars is the dominant band for independent restaurant, followed by the lowest score (1 star).

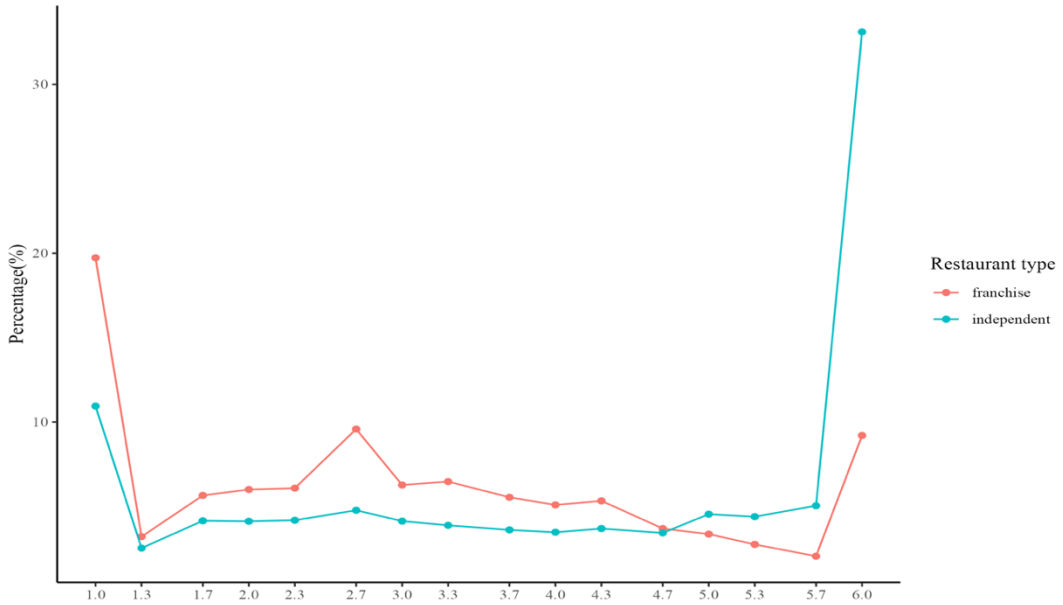

Figure 16 The rating score distribution for two types of restaurants

4.3.2 *Research design*

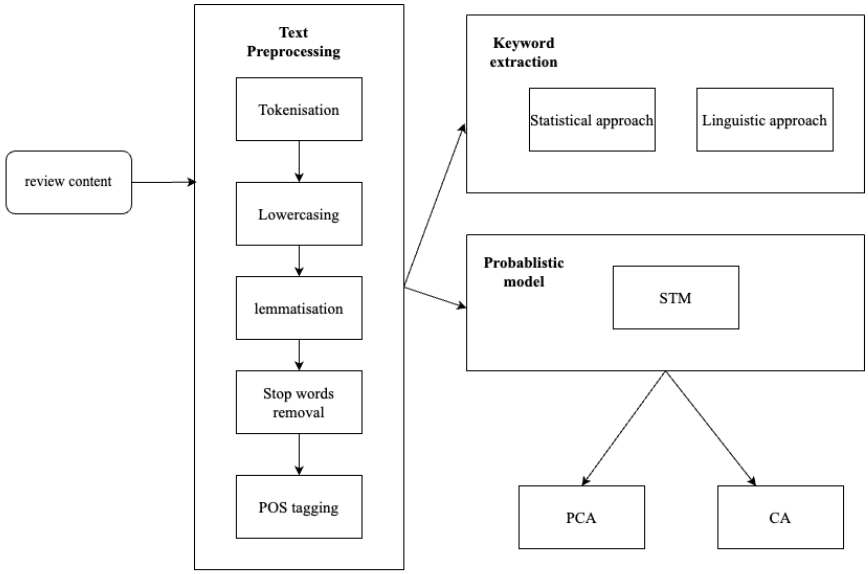

Figure 17 Research design for this study

4.3.3 *Text pre-processing*

The textual content from online reviews is written by a large number of individuals. Even it can provide us with a great deal of useful information, the user-generated reviews may contain numerous misspellings and improper word usage at first glance. Even with the finest learning algorithm, poorly prepared data cannot contribute to excellent performance without effective pre-processing (Bilalli et al., 2018). Research

indicates that only 37% of users never make typographical errors when completing online tasks (Hargittai, 2006). Moreover, textual data frequently contains special formats such as number formats and meaningless symbols. In addition, the most frequent terms, such as prepositions and articles, are not useful for text analysis. Thus, text pre-processing is required to prepare textual data for further analysis and reduce its complexity (Tirunillai and Tellis, 2014). The effective text pre-processing procedures could contribute to better model performance (Bilalli et al., 2018).

Text pre-processing enables subsequent stages to transform the text into information that is predictable and analysable. Figure 17 depicts the procedures for text pre-processing prior to text analysis. Tokenisation will be used to divide text into tokens, which may be single words or phrases. It attempts to separate sentences into a list of elements to be used as input for subsequent analysis. For example, the sentence ‘Food was very nice and delivered quickly’ will be tokenised into the following words: Food, was, 'very', ‘nice’, ‘and’, ‘delivered’, ‘quickly’. Lowercasing is a frequent stage in natural language processing (NLP) and text mining. In particular, we have utilised online evaluations from restaurant websites. Opinions are typically expressed through brief texts, which explains the prevalence of capitalised words. The main purpose of lowercasing is to standardise the text by changing all characters to lowercase. Lowercasing will therefore convert all uppercase words to lowercase words to ensure that words with different case are considered as the same token. For instance, the word "Great" will become "great."

Stemming and lemmatisation are two techniques to convert words to their root or base form. The primary objective of both approaches is to standardise various inflected or derived forms of a word so they can be regarded as the same token. Stemming is a technique to remove suffix from words to obtain their base stem, which applies a set of rules to cut common suffixes. However, stemming frequently produces inaccurate meanings because it only considers the final few letters. Lemmatisation, on the other hand, employs vocabulary and morphological analysis to transform words to the lemma, which is the standard or root form of a word. Lemmatization, unlike stemming, ensures that the resulting word is valid and has a recognised meaning. Even stemming is typically more efficient and less resource-intensive than lemmatization due to its reliance on straightforward string manipulation, lemmatisation is adopted in this study as the resulting word after lemmatisation is valid and has a recognised meaning.

Furthermore, there are a lot of words occurring in the documents frequently but are meaningless because they are used to compose complete sentences. They are called stop words and do not add help to enrich the context (Vijayarani, Ilamathi and Nithya, 2015). In contrast, their presence adds difficulties to interpreting the documents because of the high frequency stop words. For instance, 'and', 'at' and 'that' are common words contained in the text which are frequently used to compose a sentence. The system performance could be improved after reducing the text data by removing stop words. Among the meaningful words left after previous procedures, there are words that have the same meaning but different tense, such as 'is' and 'was'. Stop words are common in text, but they may not contribute significantly to its overall meaning, such as 'at', 'the', etc. For certain, stop words can be useful. It has been demonstrated that they provide meaningful insights into the personalities and behaviours of individuals (Mihalcea and Strapparava, 2009). In some cases, however, removing stop words can be advantageous, allowing the reader to focus on content words such as nouns and adjectives. To identify stop words in a text, a precompiled list of such words and an efficient lookup algorithm are typically used. In this study, stopword are selected based on SMART list (Lewis et al., 2004).

Part-of-speech (POS) tagging involves the task of assigning the correct syntactic role, such as noun, verb, etc., to each word in a given input text. Most algorithms for this task rely on supervised learning, using annotated data to learn how to assign parts of speech to new text. To extract valuable information, POS tagging will be employed to uncover the parts of speech for each word based on its definition or context. For instance, 'delicious' will be identified as adjective while 'burger' will be categorised as 'noun'. POS tagging could be considered as the last step of text pre-process and the first step in the analysis procedure as it have already assigned the linguistic role to each word.

4.3.4 Collocation analysis techniques

Statistical approach

Table 15 Contingency table of co-occurrence frequencies

$a = f(uv)$	$b = f(u\overline{v})$	$f(u*)$
$c = f(\overline{u}v)$	$d = f(\overline{uv})$	$f(\overline{u}*)$
$f(*v)$	$f(*\overline{v})$	N

The notation derived from the contingency table will be employed to illustrate the formulas for various association measures. The contingency table is introduced by Karl Pearson (1904) to analyse the cooccurrence of two words, symbolised as u, v.

The contingency table for two words is demonstrated in Table 15, which displays the observed frequencies for a word pair u, v. $f(uv)$ denotes the number of occurring times (frequency) for word pair uv in the corpus. $\overline{u}$ represents any word other than u and $\overline{v}$ bar denotes for any word except v. $*$ denotes for any word in the corpus, thus $f(u*)$ represent the total number of cooccurring times for the word pair consisting of u and any word. N stands for the total number of occurrences for any word pairs in the corpus.

The interpretation of N can differ based on the specific research objective. It generally demonstrates the total number of linguistic elements contained in the corpus, where only nouns and adjectives are included in the corpus in the collocation analysis stage.

Table 16 three types of metrics to measure the association.

	Approaches		**Metrics**
1	Simple measures of association based on frequency	Frequency-based or calculation of the joint probability	*Frequency of co-occurrence*
2	Coefficient of association strength	The degree of association between two words u and v is measured by estimating the coefficients of association strength using.	*Dice*
3	Information-Theoretic Measures	It gauges the associate strength of the words u and v in the word pair uv, which takes advantages of the concept from the information theory.	*Pointwise mutual information (PMI)*
4	Statistical Hypothesis Tests	It measures the statistical significance of the association between word u and v	*Log-likelihood Ratio; Chi-square test*

By counting the number of co-occurrences of two words, we can estimate the joint probability $P(uv)$ based on the frequency and the size of corpus.

$$P(uv) = \frac{f(uv)}{N} \tag{28}$$

For PMI: It is demonstrated by Pecina (2010) that the PMI is the most effective method to measure the association for identifying collocations based on the empirically comparative analysis of various association measures. PMI is alculated as below:

$$PMI = \log_2 \frac{P(uv)}{P(u*) * P(*v)} \tag{29}$$

Nevertheless, PMI could face difficulties when working with sparse data, where specific word combinations occur infrequently but have valuable semantic meaning. In addition, pointwise mutual information may not produce accurate results for collocations involving word pairings with low frequency. Consequently, despite the fact that this metric is useful in many situations, its limitations must be considered when applying it to particular datasets.

This association measure (D) is calculated according to the formula.

$$D = \frac{2f(uv)}{f(u*) + f(*v)} \tag{30}$$

This coefficient stands out as one of the most commonly employed measures of association for detecting collocations. Its performance has been observed to exceed that of other association measures. For instance, the Dice metric outperformed the other metrics when extracting collocations from the research of (Rychlý, 2018). Therefore, it is the preferred option when attempting to identify meaningful word combinations or collocations in a variety of textual contexts.

This approach aims to assess the significance association between two words u and v within the word pair uv. It can quantitatively evaluate the amount of evidence that the null hypothesis (independence between words u and v are or homogeneity within the contingency table) is rejected using the observed data in the contingency table. It can distinguish meaningful word associations or collocations from random occurrences by evaluating the evidence against the null hypotheses. Two methods are selected from this approach: Log-likelihood ratio and Chi-square test.

The log-likelihood ratio is calculated as the following formula:

$$LLR = 2\sum_{i,j} O_{ij}\, log\, \frac{O_{ij}}{E_{ij}} \tag{31}$$

where O_{ij} represents the observed frequencies in the contingency table with $i * j$ dimensions as we focus on the combinations of two words. E_{ij} are the expected frequencies under the null hypothesis for cell ij. In this case, they would be calculated as:

$$LLR = 2\left[f(uv) * \log\left(\frac{f(uv)}{E_{uv}}\right) + f(u\bar{v}) * \log\left(\frac{f(u\bar{v})}{E_{u\bar{v}}}\right) + f(\bar{u}v) * \log\left(\frac{f(\bar{u}v)}{E_{\bar{u}v}}\right) + f(\overline{uv}) * \log\left(\frac{f(\overline{uv})}{E_{\overline{uv}}}\right)\right] \tag{32}$$

$$E_{element1,element2} = \frac{f(element1) * f(element2)}{N} \quad (33)$$

where element1could be u and $\bar{u}$ and element 2 could be v and $\bar{v}$.

Chi-square test is operated under the normal distribution assumption. The asymptotic theory could be utilised because of the large sample of linguistic data. The statistical tests mentioned before work with finite data samples. Nevertheless, this method test statistics for such infinitely large sample sizes under the asymptotic theory.

The Pearson's χ^2 test is employed to test the independence of data observed in the contingency table using the below formula.

$$\chi^2 = \sum_{u,v} \frac{(f_{uv} - \widehat{f_{u,v}})^2}{f_{u,v}} \quad (34)$$

Linguistic approach

Collocations are characterised as distinct and frequently occurring combinations (Gelbukh et al., 2004) of two linguistic elements. They are frequently found together due to a direct syntactic relationship. Nevertheless, their co-occurrence in the corpus cannot be explained solely by grammatical rules (Ebrahimi and Toosi, 2013). In the collocation analysis, the potential word combinations are identified by several approaches. Two types of approaches including linguistic approach and statistical approach are employed to extract the collocation and identify the keywords.

When calculating association measures, it is customary to consider the frequencies of occurrence for each word and the co-occurrence for the word pair. In order to discover the collocation or extract the keywords, linguistic approaches are frequently employed utilizing morphological or syntactic details. For example, with the help of POS tags after the text preprocessing, nouns and adjectives could be identified because of their more intensive information contained.

Diverse theoretical frameworks hold diverse assumptions regarding the particular complexities of syntactic structure. Nevertheless, all forms of dependency grammar share a fundamental concept which is syntactic structure consists predominantly of words connected by binary, asymmetrical relationships known as dependency relations or dependencies. These relationships exist between a word with

syntactic subordination, known as the *dependent*, and another word, known as the *head*, on which it depends for its syntactic function.

This concept is depicted visually in Figure 18, which illustrates the dependency structure of one sentence from customer review. The dependency relationships could be observed in this structure, indicated by arrows that extend from the head to the dependent terms (Nivre, 2010). Each arrow is labelled to signify the nature of dependency it represents, which is called *dependency type*.

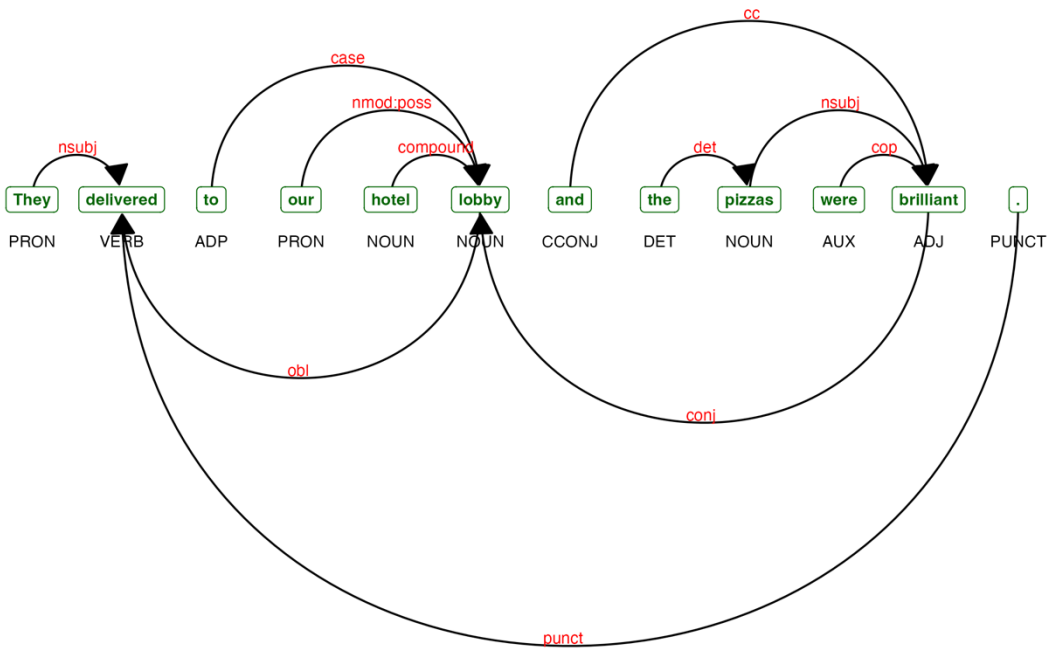

Figure 18 An example of dependency structure

The inventory of dependency relations used in this study is derived from the research by De Marneffe et al. (2014). Among the over 40 dependency relations, 'nsubj' plays a vital role in this study as we focused on the adjectives and nouns pairs.

4.3.5 Structural topic model

Latent Dirichlet Allocation (LDA) is the most common topic modelling approach following a generative probabilistic model based on Bayesian Inference (Blei et al., 2003). The core assumption of LDA is that each document is generated by several '*topics'*, which are unobservable (i.e., 'latent'). Each document is defined as a mixture over topics (admixture model), with each topic having a probability between 0 and 1 and summing up to 1. This distribution of topics belonging to document is defined as

document-topic distribution or theta, which is drawn from a Dirichlet prior. Each word is considered as a vector with a probability denoting a topic. The distribution of words denoting a topic is defined as topic-word distribution, which is drawn from another Dirichlet distribution.

Even though LDA has been applied to analyse unstructured data in several fields, its drawback is the lack of exogenous covariates in the process, which could negatively affect the reliability of the topic solution (Korfiatis, Stamolampros, Kourouthanassis, and Sagiadinos, 2019). The structural topic modelling approach (STM) proposed by Roberts et al. (2014) could address this drawback. Compared with LDA, STM allows one to introduce document-level covariates (the characteristics of reviews in this study) to influence the topic proportions. This can show how topic prevalence changes with document-level covariates (Roberts, Stewart and Airoldi, 2016). Figure 19 represents the STM process for the review text from customer reviews, using the plate notation. In this study, each review text is indexed by $r \in \{1,2, \dots R\}$. K is assumed to be the number of topics. Each word in the vocabulary is indexed by $v \in (1, \dots, V)$. STM contains three components: i) topic prevalence model, ii) the core language model and iii) topical content model. The topic prevalence model and the topical content model control how covariates influence document-topic proportions and topic-word distributions respectively. The language model controls actual words production in each review.

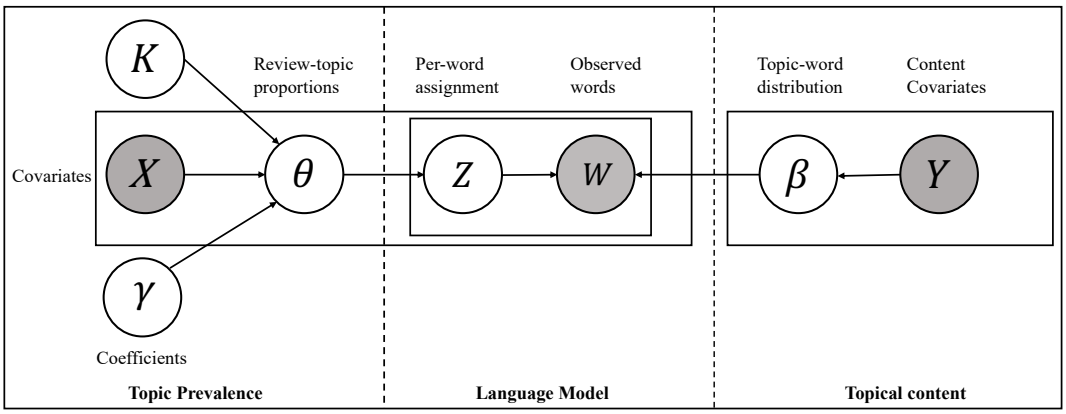

Figure 19 A graphical demonstration of STM process (adopted by Roberts et al., 2016)

The generative process is summarised as follows:

- Draw document-level attention to topics from a logistic-normal generalised linear model accounting for several review-level covariates X_r.

$$\underset{\theta_r}{\rightarrow} |X_{r\gamma} \sim LogisticNormal\ (\mu = X_{r\gamma}, \Sigma) \tag{35}$$

where Σ is a $(K-1) \times (K-1)$ covariance matrix. γ represents a $p \times (K-1)$ matrix of coefficients drawn from a normal distribution for each k ($k = 1, \ldots, K-1$) shown as below:

$$\gamma_k \sim \text{Normal}(0, \sigma_k^2) \tag{36}$$

$$\sigma_k^2 \sim \text{InverseGamma(a, b)} \tag{37}$$

where a and b are fixed hyperparameters (Roberts et al., 2016).

- The document-specific distribution($\beta_{r,k}$) over words to which represent each topic (k) is formed of the general word distribution m, the deviations of distributions across topics $\kappa_k^{(t)}$, the group deviation of covariate $\kappa_{yd}^{(c)}$, and the interaction between the two deviations $\kappa_{y_r,k}^{(i)}$.

$$\beta_{r,k} \propto \exp\left(m + \kappa_k^{(t)} + \kappa_{y_r}^{(c)} + \kappa_{y_r,k}^{(i)}\right) \tag{38}$$

The following steps assume the core model of STM.

- First, draw $z_{r,n}$ from the review-specific distribution over topic:

$$z_{r,n}|\underset{\theta_r}{\rightarrow} \sim \text{Multinomial}(\theta_r) \tag{39}$$

- Then, an overserved word $w_{r,n}$ is chosen from the distribution over topical content parameters $\beta_{r,k,v}$, which represents the review-specific distribution over topics.

$$w_{r,n}|Z_{r,n}, \beta_{d,k} = z_{r,n} \sim \text{Multinomial}(\beta_{r,k} = z_{r,n}) \tag{40}$$

Following Roberts et al. (2016), we adopt the FREX metric to measure topic quality through a combination of frequency and exclusivity of word to topics.

$$FREX_{k,v} = \left(\frac{\omega}{ECDF\left(\frac{\beta_{k,v}}{\sum_{j=1}^{K}\beta_{j,v}}\right)} + \frac{1-\omega}{ECDF(\beta_{k,v})}\right)^{-1} \tag{41}$$

where ω is the weight to favour exclusivity (we set to 0.7) and ECDF is the empirical CDF.

4.3.6 Model estimation

We use the 'stm' package in R to extract the latent topics. As mentioned, we include the characteristics of the customer reviews as covariates in the model, to influence the topic prevalence. The primary variables of interest are the review rating by the customer, representing customer satisfaction, and a dummy variable capturing the type of restaurant under review (*franchised* or *independent*).

To prepare the corpus, review comments are pre-processed based on the following steps: i) word tokenisation (separating sentences into a list of tokens), ii) stop words removal, including language stop words based on the SMART list (Lewis et al., 2004) and subject-specific stop words (e.g., names of restaurants and food vocabulary), iii) selecting only nouns, adjectives, and adverbs as these words indicate consumers' assessment about products/services. This step is carried out using POS (part-of-speech) tagging in addition to lemmatisation to get their base forms. We excluded words whose frequency appeared in less than 1% of the total number of reviews. This process reduced the number of reviews to 538,623.

To explicit the optimal number of topics (K), we adopt four criteria to evaluate the model outcomes: semantic coherence, exclusivity, held-out likelihood and residuals. Semantic coherence is a metric developed by Mimno et al. (2011), which is maximised when the most probable words within one topic co-occur frequently together. Exclusivity refers to the degree to which topics are distinct from each other. Held-out likelihood captures how much of the variability is explained by the topic solution. Residual checks the overdispersion of the variance of the multinomial in the generating process of STM (Taddy, 2012). Based on the intrinsic nature of the reviews, we experiment with values of K (number of topics) between 6 and 17 and evaluate different criteria of the structural topic model.

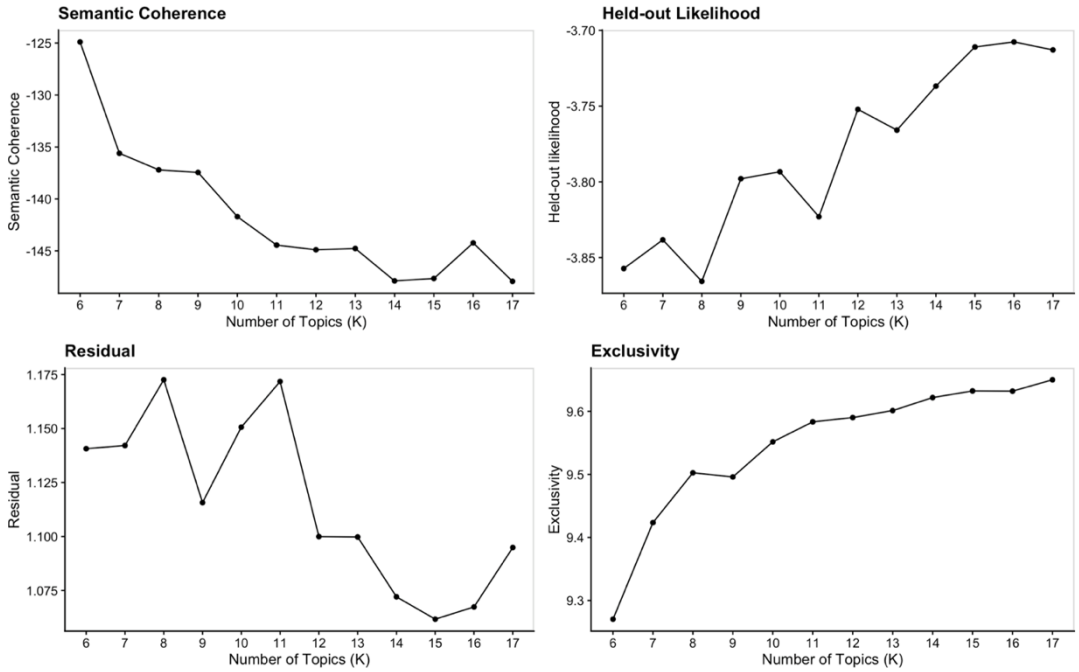

Figure 20 Diagnostic values for topic solution with different K numbers

The plot demonstrates four diagnostic values we adopt to evaluate the model with different number of topics (K). We believe that the best solution is achieved when K=15, since this is when the residuals are minimised while held-out likelihood and exclusivity take highly satisfactory values.

4.4 Results

4.4.1 Text statistics

After the final phase of text pre-processing, each word in the provided text is assigned its POS tag. Each term in the review text is marked as a specific portion of speech based on its definition and context. There are various POS tags, such as noun, verb, adjective, etc. In the OFD service context, nouns and adjectives are vital tags customers express their opinions and attitudes towards the food and services while other tags (such as verbs) could be helpful in another context.

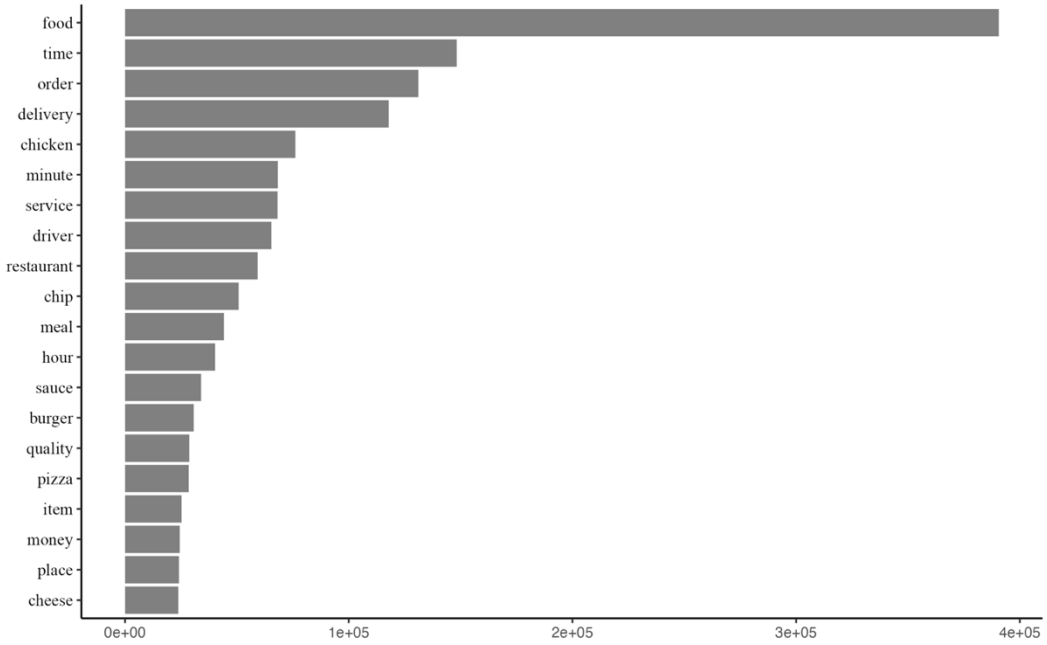

Figure 21 Top nouns across the corpus

Based on the frequency, both nouns and adjectives are selected and displayed in Figure 21 and Figure 22 in descending order. Figure 21 demonstrates the top 20 most frequently occurring nouns in the corpus-level. Each bar in the figure represents a specific word and the height of each bar means the occurring times of each word across the whole corpus. The x-axis represents the frequency count to the corresponding word.

As shown in the figure, the dominant word is ‘food’, which occurs almost 0.4 million times across the whole corpus. Three words ‘time’, ‘order’, and ‘delivery’ with high occurrence are followed by ‘food’, which amount to over 0.1 million times. The most frequent words could be roughly categorised into two groups. One covers the food types, which refer to the detailed food name, such as ‘chicken’, ‘chip’, and ‘burger’. The other group contains words describing their experiences, for instance, ‘time’, ‘delivery’, ‘service’, etc.

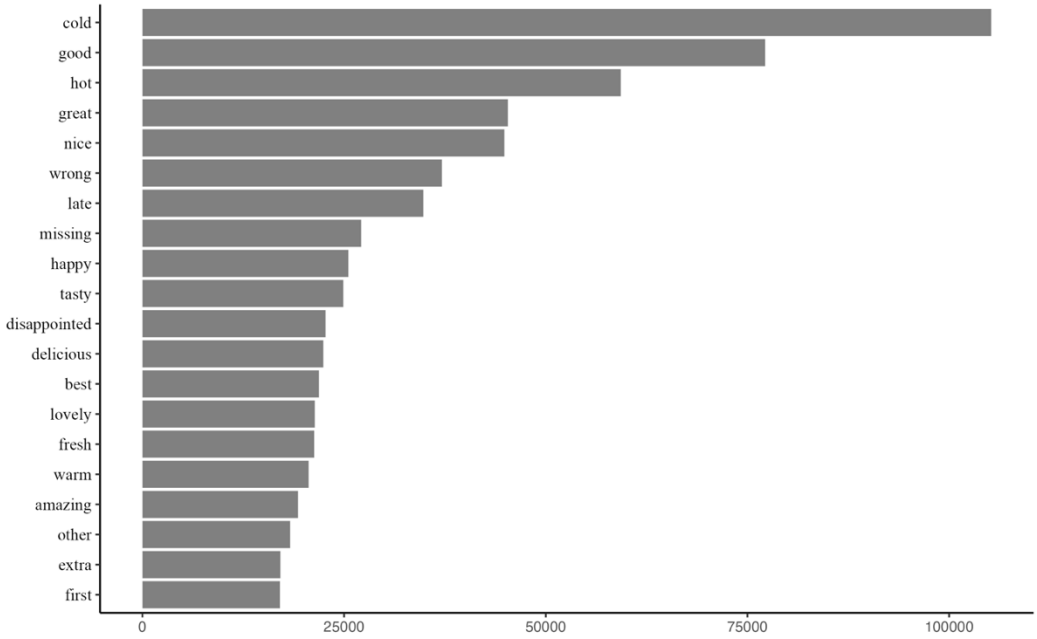

Figure 22 Top adjectives across the corpus

Apart from nouns, Figure 22 illustrates the most occurring adjectives (top 20) across the whole corpus. The most prevalent word is ‘cold’ occurring over 0.1 million times, followed by ‘good’, which occurs around 75 thousand times. Some words including ‘hot’, ‘great’, and ‘nice’ also have high frequency in the corpus level. Some adjectives are more likely to describe ‘food’ such as ‘hot’ and ‘delicious. However, some adjectives could not be inferred arbitrarily. For examples, ‘great’ and ‘nice’ could be used to describe the food quality as well as the delivery service. Therefore, the adjective-noun pair would be more helpful to understand the whole corpus since adjectives occurring with nouns are able to provide more information about the nouns (Yule, 2022).

4.4.2 Collocation analysis

In accordance with the methodology described in previous section, adjective-noun pairs have been extracted by utilising the dependency relationship using the linguistic approach. These pairs are able to encapsulate the most frequent and relevant expression of customer feelings regarding the food and delivery service. Table 17 ranks the most frequent adjective-noun pairs based on their occurrence frequency. Each pair is a condensed representation of consumer perceptions regarding their

experiences. The ranking is determined by the frequency with which each combination appears in the textual content of reviews.

As shown in the table, it is not surprisingly to be observed that adjectives such as "cold," "good," and "hot" are frequently paired with the noun "food" as 'food' is the most frequently mentioned word in reviews. They emphasise the significance of food quality and temperature in terms of consumer evaluations. Furthermore, the occurrence of phrases such as "late food" and "late delivery" indicates customers' dissatisfaction towards the delivery service, especially the delivery speed.

Table 17 Top adjective-noun pairs

Rank	Adjective-Noun	Frequency	Proportion
1	cold food	26816	7.67%
2	good food	11277	3.22%
3	hot food	9926	2.84%
4	nice food	7488	2.14%
5	late food	7484	2.14%
6	great food	7122	2.04%
7	delicious food	4888	1.4%
8	amazing food	4695	1.34%
9	lovely food	4639	1.33%
10	wrong order	4493	1.28%
11	warm food	4164	1.19%
12	cold chip	3844	1.1%
13	cold stone	3743	1.07%
14	ok food	2931	0.84%
15	tasty food	2745	0.78%
16	excellent food	2231	0.64%
17	late order	2021	0.58%
18	fresh food	1987	0.57%
19	friendly driver	1899	0.54%
20	late delivery	1584	0.45%
21	soggy chip	1431	0.41%
22	rude driver	1421	0.41%
23	dry chicken	1343	0.38%
24	good quality	1219	0.35%
25	correct order	1150	0.33%
26	nice driver	1114	0.32%
27	polite driver	1086	0.31%
28	quick delivery	1040	0.3%
29	fine food	1038	0.3%
30	good service	1021	0.29%

Customers are concerned about a variety of factors, such as the driver's demeanour and the rate of delivery, as illustrated by a number of pairings. There are

several phrases such as "friendly driver," "quick delivery," and "politer driver." These pairings illustrate the influence of the delivery service on the overall perception of meal delivery service.

However, the information contained in the adjective-noun pairs is limited to its characteristics. As we shown in the noun frequency plot, food is the most dominant noun in the corpus, approximately 4 times higher frequency than the second prevalent noun. Because of the dominance of ‘food’, the adjective-noun pair is dominated by the adjective related to the word ‘food’. Therefore, further analysis would follow the procedures to identify the word collocations. First, the most occurring nouns are selected based on the frequency of occurrence across the whole corpus. Then 5 metrics mentioned are calculate within the same sentence to identify the word collocations. The results are summarised in below tables.

The following four tables depict the results of collocation analysis for the top 4 most occurring nouns (‘food’, ‘time’, ‘order’, ‘delivery’) in the corpus. The values of various metrics are shown in the columns, including Frequency (Freq), Mutual Information (MI), Dice Coefficient (Dice), Log-Likelihood Ratio (LL), and Chi-Square (Chi). After sorting the values of each metrics from high to low and selected top 10 values, the corresponding terms are also displayed alongside the value. Collectively, the metrics provide a multifaceted evaluation of associations, allowing the identification of terms with varying degrees of connection to the core terms.

It could be inferred from four tables that in our case. The two metrics measuring statistical significance of the association perform better than other metrics as the terms are able to provide more meaningful information and easier for us to interpret. They avoid the influence of those words with high frequencies on the collocations. Specifically, we are able to explore the related words to every word we are interest in. It creates more efficiency instead of identifying customers’ concerning aspects from the most occurring adjective-noun pairs. For instance, from Table 17 the term ‘order’ is only mentioned 3 times in the top 30 pairs. However, from

Table 20, ‘order’ is tested to be associated with ‘wrong’, ‘item’, ‘correct’, ‘missing’, etc.

Subsequently, the top 100 occurring 100 nouns are selected and LL is utilised to identify the associated terms of each noun as well its LL value. We have removed the nouns which depict the specific food name (such as ‘burger’, ‘chicken’, ‘curry’,

etc.) Based on the LL value, we construct a word collocation network to illustrate how these words are connected to each other.

Table 18 Metrics to identify the collocation of “food”

Freq-terms	Freq	MI-terms	MI	Dice-Terms	Dice	LL-Terms	LL	Chi-Terms	Chi-square
food	360121	food	1.32	food	1.00	cold	60999.03	food	1352360.00
cold	62283	agency	1.24	cold	0.27	hot	30922.08	cold	85790.99
time	53609	clod	1.23	time	0.21	quality	16767.03	hot	40683.44
delivery	41724	rucksack	1.14	delivery	0.18	great	14239.23	time	22283.18
hot	35997	stun	1.12	hot	0.17	good	12538.40	quality	20751.46
good	35829	frm	1.10	good	0.16	late	11497.67	good	19962.24
order	30322	quilty	1.10	great	0.13	warm	9887.48	great	19475.30
great	28084	coma	1.06	hour	0.13	hour	9370.68	delivery	15166.45
hour	26401	okayish	1.06	order	0.12	lovely	7788.96	late	14564.50
driver	23798	enroute	1.06	driver	0.11	amazing	6944.83	hour	13474.48

Table 19 Metrics to identify the collocation of “time”

Freq-terms	Freq	MI-terms	MI	Dice-Terms	Dice	LL-Terms	LL	Chi-	Chi-square
time	155924	time	2.16	time	1.00	first	70348.37	time	1352360.00
food	53609	approximate	2.16	delivery	0.24	delivery	27048.76	first	132197.62
delivery	32413	frame	2.13	first	0.24	second	21597.92	delivery	123278.19
order	23539	allocated	2.12	food	0.21	last	20963.64	order	110436.31
first	21616	predicted	2.11	order	0.16	long	14018.22	second	39818.19
cold	11903	allocate	2.10	last	0.11	next	12642.97	last	37542.53
good	11811	frist	2.08	hour	0.10	few	5983.35	long	25013.66
hour	10819	forth	2.07	hot	0.10	2nd	5821.57	next	22401.94
hot	10777	changing	2.07	good	0.10	food	5174.98	2nd	10601.63
restaurant	10296	unprecedented	2.07	restaurant	0.09	1st	5033.61	cold	9777.16

Table 20 Metrics to identify the collocation of “order”

Freq-terms	Freq	MI-terms	MI	Dice-Terms	Dice	LL-Terms	LL	Chi-Terms	Chi-square
order	142289	order	2.25	order	1.00	wrong	34714.93	order	1352360.00
food	30322	placing	2.21	wrong	0.20	item	10163.65	time	502182.59
time	23539	misse	2.16	time	0.16	correct	9654.27	wrong	69881.45
wrong	18124	incomplete	2.08	hour	0.13	missing	8450.15	delivery	33657.82
delivery	14170	placement	1.93	food	0.12	part	7823.46	item	19025.43
hour	12612	protest	1.85	item	0.11	hour	6223.91	hour	18519.48
cold	10885	brownd	1.80	delivery	0.11	incorrect	3908.48	correct	17556.82
restaurant	10775	duplicate	1.78	restaurant	0.11	time	3560.67	missing	15160.44
driver	9819	odef	1.77	missing	0.10	restaurant	2875.92	part	13861.26
item	9583	odeff	1.75	driver	0.09	first	2781.40	cold	12843.86

Table 21 Metrics to identify the collocation of “delivery”

Freq-terms	Freq	MI-terms	MI	Dice-Terms	Dice	LL-Terms	LL	Chi-Terms	Chi-square
delivery	113846	delivery	2.47	delivery	1.00	driver	42699.03	delivery	1352360.00
food	41724	-contact	2.38	driver	0.27	time	27048.76	driver	131302.55
time	32413	fee	2.31	time	0.24	quick	14324.80	quick	29334.14
driver	23945	contactless	2.31	food	0.18	fast	13800.59	fast	28727.14
order	14170	-free	2.25	hour	0.12	friendly	11536.96	cold	27991.31
service	10960	fastest	2.25	service	0.12	man	10597.97	service	24858.11
hour	10741	superfast	2.18	order	0.11	polite	10535.12	hour	23291.57
good	9280	plesent	2.16	minute	0.10	guy	9015.61	friendly	22519.00
cold	8171	speedy	2.14	late	0.10	charge	6797.86	man	21508.52
great	8035	charge	2.10	good	0.10	food	6013.30	polite	20606.06

The generated word network is a visual representation that encapsulates the of associations among various terms within the corpus. In this constructed network, each term serves as a node, while the edges connecting them are weighted by their respective log likelihood values. The width of the edges acts as a visual indicator of the strength and likelihood of relationships between pairs of words. As shown in Figure 23 there are several clusters that show various aspects of customer evaluations, such as the customer service, the delivery driver, the order issues, etc.

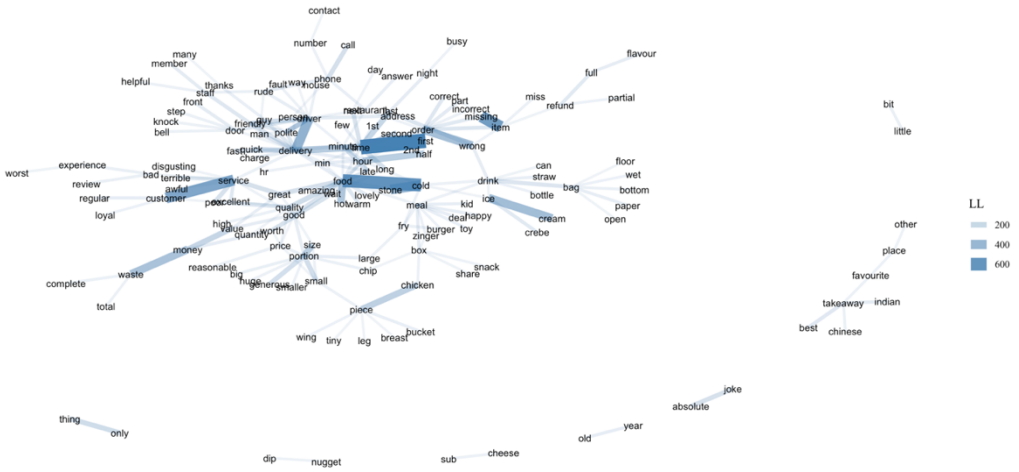

Figure 23 A Network showing the collocation

4.4.3 Topic solution

Table 22 presents an overview of the 15-topic solution. An LDA process produces two matrices: the document-topic proportion matrix (theta matrix), and the word-topic distribution matrix. The third column shows the corpus-level topic proportions (the distributions of each topic across the whole corpus), calculated as the average of each topic's proportion within all reviews. We manually label each topic using the top loading words utilising the FREX score.

Table 22 Labels, distribution and top 7 loading words in the topic solution.

#	Topic Label	Prop (%)	Top 7 FREX words
1	Cold food	15.58	cold, late, soggy, food, hour, stone, warm
2	Repeat purchase	7.62	first, last, time, second, long, next, twice
3	Customer rumination	2.76	money, back, worth, waste, people, lot, bin
4	Order issues (incorrectness)	4.99	wrong, refund, part, right, joke, note, full
5	Food quality (critique)	5.66	dry, hard, never, worst, awful, disappointed, bad
6	Delivery service (praise)	8.17	great, friendly, excellent, lovely, quick, fast, early
7	Delivery service (critique)	8.22	door, driver, address, phone, rude, house, delivery
8	Value for money	6.10	portion, good, small, price, quality, size, enough
9	Order issues (incompleteness)	7.81	item, missing, order, miss, issue, free, sweet
10	Takeaway experience	5.75	best, always, ever, amazing, takeaway, place, highly
11	Packaging issues	3.82	drink, bag, ice, half, fine, large, top
12	Meal deal	6.09	meal, happy, extra, box, instead, double, one
13	Customer service	3.83	customer, poor, service, terrible, staff, food, issue
14	Food quality (praise)	9.98	nice, tasty, fresh, hot, spicy, delicious, food
15	Delivery time	3.63	later, min, busy, minute, shame, shop, food

According to Table 22, topics related to food quality (Topics 1, 5, 14) and delivery service (Topics 6,7 and 15), corresponding to the two critical aspects in the OFD service supply chain (food production and service delivery), account for a combined proportion of 51.24%. As demonstrated by the top 7 words, the food quality topics discuss various features regarding food quality that could affect customer

satisfaction, including food temperature, freshness and taste. The aspects mentioned concerning delivery include waiting time, drivers' politeness and the problem of finding house. Moreover, three topics seem to refer to the order processing stage. They identify issues that arise during the preparation of orders, including the absence of part of the order, the delivery of an incorrect order, and issues with the packaging that result in customers not receiving their food in the intended state. The remaining topics describe the customer's experience with the process, and their future purchase intention. For each topic, the word clouds showing the top words are shown in Table A1 (Appendix A).

To show how these topics distribute differently in franchised and independent restaurants, we re-calculated the topic proportions for the two types of restaurants separately and conducted Mann-Whitney U-tests to detect statistically significant differences between the two. As shown in Table 23 , for all topics the differences in topic proportions are highly statistically significant. This suggests that when customers provide reviews to the two types of restaurants, they tend to focus on different aspects of their experience.

Table 23 Non-parametric Mann-Whitney U tests between the two types of restaurants

#	Topic Label	Franchised Proportion (%)	Independent Proportion (%)	W_value (*10^{11})
1	Cold food	18.26	12.41	4.55^{***}
2	Repeat purchase	7.25	8.07	2.69^{***}
3	Customer rumination	2.64	2.91	2.87^{***}
4	Order issues (incorrectness)	6.80	2.86	6.11^{***}
5	Food quality (critique)	5.52	5.83	3.57^{***}
6	Delivery service (praise)	5.46	11.38	1.81^{***}
7	Delivery service (critique)	9.67	6.50	5.14^{***}
8	Value for money	4.20	8.35	1.12^{***}
9	Order issues (incompleteness)	11.01	4.03	6.19^{***}
10	Takeaway experience	2.82	9.22	0.99^{***}

#	Topic Label	Franchised Proportion (%)	Independent Proportion (%)	W_value (*10^{11})
11	Packaging issues	5.72	1.57	6.34***
12	Meal deal	8.00	3.82	5.95***
13	Customer service	4.03	3.59	4.10***
14	Food quality (praise)	5.48	15.30	1.06***
15	Delivery time	3.16	4.18	2.87***

Notes: *Reported proportions are averages across reviews. ***: p-value < 0.001*

As mentioned, the biggest advantage of the given topic modelling method is to allow document-level covariates to influence the topic solution. As such, we can measure systematic changes in topic prevalence over the document-level covariates. Specifically, we obtain estimates of how review ratings or restaurant type affect topic prevalence, when holding the other variable constant (Roberts et al., 2014). These estimate coefficients are derived from a fractional logistic regression model as follows:

$$\theta_r = \beta_1 RestType_r + \beta_2 RevRating_r + \beta_3 RevRating_r * RestType_r + \varepsilon_r \quad (42)$$

where θ_r is the topic proportion in review r, $RestType_r$ represents the type of restaurant r, $RevRating_r$ is the numerical rating. $RevRating_r * RestType_r$ represents the interaction term.

By averaging the estimation of the topic proportion across the levels of the categorical variables we are able to estimate comparative estimations across categories and more specifically for low (1 star) and high rating (6 stars – as advertised by the platform), as well as for franchised and independent businesses.

Figure 24 The changes in the expected topic proportion from lower rating to higher rating for all restaurants

Figure 24 displays the marginal effects of review ratings on each topic proportion when the other variable is held at its sample median (Roberts, Stewart, and Tingley, 2019). The dotted line represents the zero effect. The length of a topic's distance from the dotted line indicates the degree to which its proportion changes as the review rating increases (or decreases) from the lowest (1) to the highest rating (6) (or vice versa). As shown, when we shift from a rating of 1 to a rating of 6, all predicted proportions of topics reflecting praise toward an aspect increase, while those expressing critique decrease. The topic whose proportion is affected the most is delivery service (praise): On average, if the review rating changes from 1 to 6, the proportion of delivery service (praise) will increase. In other words, praise for delivery service is much more likely to appear in reviews with a higher rating. This topic is followed by Food quality (praise). On the other hand, cold food demonstrates the most significant decrease in the topic proportion within a review that shifts from the lowest to the highest rating. Food quality (critique) follows suit while the order incompleteness shows the third largest decrease in topic proportion.

4.4.4 Comparison between franchised and independent restaurants

Figure 25 displays the marginal effects of restaurant type on each topic proportion. The dotted line represents the zero effect. The length of a topic's distance from the dotted line indicates how its proportion would change if the restaurant under review

changed from being independent to being franchised. In other words, the plot shows the extent to which the expected proportion of each topic differs for the two types of restaurants. On the left-hand side of the dotted line, ten topics are more popular in reviews of independent restaurants, while those five on the right side are more dominant in reviews of franchised restaurants.

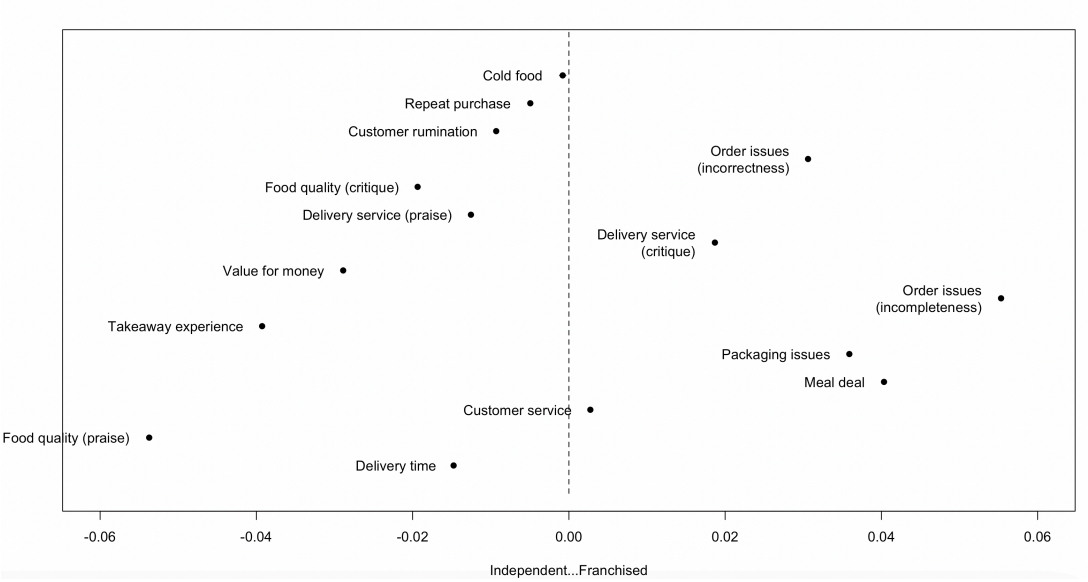

Figure 25 The changes in the expected topic proportion if the restaurant changed from being independent to being franchised

As shown, the topics ‘order issues (incompleteness)’ and ‘Order issues (incorrectness)’ are 5.93% and 3% more likely to appear in a review of a franchised restaurant (for a review rating at its sample mean). Both are related to order processing activities. This would suggest that franchised restaurants underperform independent restaurants in this respect, which might result from the high volume of orders that franchised restaurants might receive. In addition, on average, customers of franchised restaurants are more critical regarding delivery service. On the other hand, an independent restaurant of mean rating (3.86) is more likely to be praised for its food quality and delivery service, compared to a franchised restaurant. This might result from the uniqueness of the food taste, and a friendlier interaction between drivers and customers. However, a customer of an independent restaurant is also more likely to complain about the food quality and delivery time. These might be due to the fact that processes in independent restaurants are relatively less likely to be highly automated and standardised, so they are more prone to unexpected deviations from the norm (especially during busy times), leading to delays, over- or under-cooked food.

Nevertheless, since praising the takeaway experience with independent restaurants (with superlatives such as ‘best’ and ‘amazing’) is also considerably more likely, it can be concluded that, on balance, the review of an independent restaurant of an ‘average’ rating is relatively more positive.

To demonstrate the internal relationship among the topics for two types of restaurants, we performed a PCA (Principal Component Analysis) in the two sub-samples, using the prevalence scores of the 15 topics extracted from the STM. The first subset includes 247,241 customer reviews form franchised restaurants of 31 brands. The expected topic proportions (θ) are aggregated at the restaurant level. The second subset includes 291,382 customer reviews of independent restaurants. Similarly, topic memberships were aggregated at the restaurant level. Figure 26 shows how the topics are correlated in the reviews of the two types of restaurants. The first two principal components explained 59% of the variance for franchised restaurants and 56.2% for independent ones. As shown, for independent restaurants, the topics are relatively more polarised, suggesting that reviews for these restaurants carry stronger valence (either very positive or very negative. Thus, a customer satisfied (dissatisfied) with their experience with an independent restaurant is likely to touch on many positive (negative) aspects in their textual review.

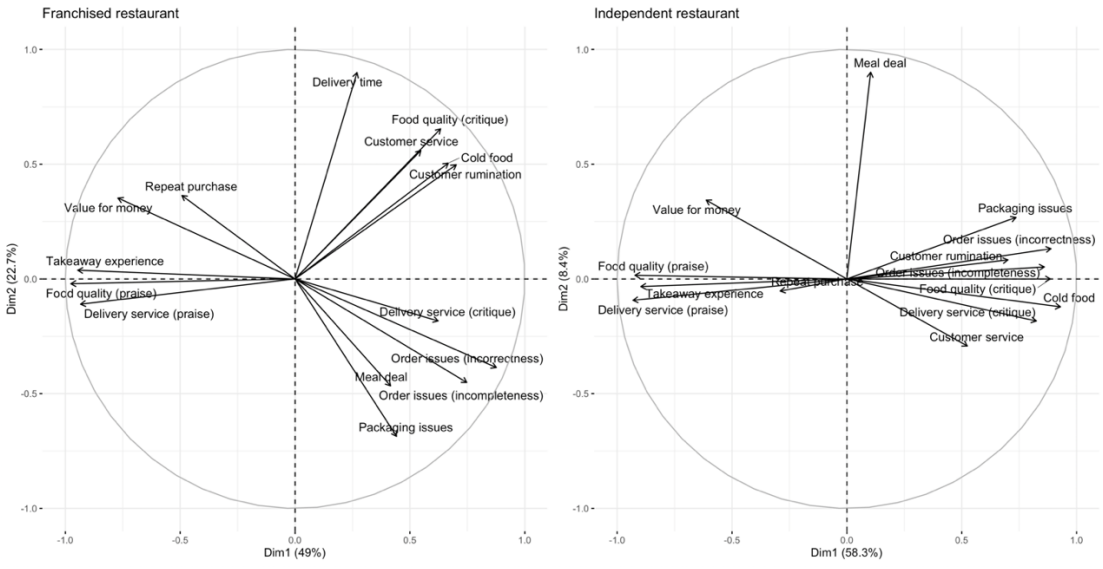

Figure 26 Principal component analysis: topic correlation plots for franchised versus independent restaurants

4.4.5 The interaction of the restaurant type

To demonstrate the interaction between the restaurant type and review score, we display the different marginal effects of two types of restaurants on the topic distribution between extremely negative reviews (left-hand side) and extremely positive reviews (right-hand side). As shown in the figure, several topics indicate similar changes when we shift from independent restaurant to franchised restaurant between positive reviews and negative reviews, such as 'customer rumination', 'value for money', order-related issues, etc. Nevertheless, the changes of proportions of several topics for two types of restaurants diversify between extreme negative group and positive group.

Figure 27 The interaction between rating score and restaurant type

In the negative rating group, several topics revealing customer dissatisfaction are more prevalent in reviews from independent restaurants including 'cold food', 'delivery service critique', 'food quality critique' and 'delivery time' while in the positive rating group, these topics have higher proportions in reviews from franchised restaurant. No matter in negative rating group or positive rating group, topics showing the compliment from customers are more prevalent in reviews from independent restaurants, such as the topics 'food quality praise' and 'delivery service praise'.

4.4.6 Digging into franchised restaurants

Except from the comparison between franchised and independent restaurant, we pay attention to the heterogeneity across 31 brands regarding customer evaluation on the service quality provided by franchisees. Table 24 summarises the average

rating (AR) and the average expected topic proportions for each topic at the brand level and the number of franchisees on JustEat, displayed from the highest to the lowest. Based on the franchisee amount, the table only shows the top 7 and the bottom 7 brands. As shown, Subway has the most franchisees in the UK (1283), followed by McDonald's (867) and KFC (681) between the period of four years. We employed one-way ANOVA to compare the difference between the brands for the 15 topics' proportions. The last row displays the F statistic value and the corresponding significance.

Table 24 Expected topic proportion across all brands from franchised restaurants

Brand	AR	FA	Topic 1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	Topic10	Topic11	Topic12	Topic13	Topic14	Topic15
Subway	3.31	1283	0.16	0.07	0.03	0.07	0.05	0.07	0.10	0.05	0.11	0.03	0.05	0.08	0.04	0.06	0.03
McDonald's®	2.82	867	0.20	0.07	0.03	0.07	0.05	0.04	0.10	0.04	0.11	0.03	0.07	0.08	0.04	0.05	0.03
KFC	3.02	681	0.18	0.07	0.03	0.07	0.06	0.05	0.10	0.04	0.12	0.03	0.05	0.08	0.04	0.05	0.03
Greggs	3.77	668	0.14	0.07	0.02	0.06	0.05	0.09	0.10	0.05	0.11	0.04	0.05	0.07	0.04	0.08	0.03
Burger King	3.02	244	0.19	0.07	0.03	0.07	0.06	0.06	0.10	0.04	0.11	0.03	0.05	0.08	0.04	0.06	0.03
STARBUCKS®	3.59	237	0.13	0.07	0.02	0.07	0.05	0.08	0.10	0.05	0.10	0.03	0.09	0.07	0.04	0.07	0.03
Pret A Manger	3.63	136	0.14	0.07	0.02	0.07	0.05	0.07	0.10	0.05	0.12	0.03	0.06	0.07	0.04	0.07	0.03
…																	
…																	

Brand	AR	FA	Topic 1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	Topic10	Topic11	Topic12	Topic13	Topic14	Topic15
Heavenly Desserts	3.65	21	0.15	0.08	0.03	0.06	0.05	0.09	0.09	0.07	0.09	0.04	0.05	0.07	0.04	0.08	0.03
McCain SureCrisp	3.98	18	0.14	0.08	0.02	0.05	0.05	0.10	0.09	0.06	0.08	0.05	0.04	0.07	0.04	0.10	0.03
Chiquito	3.73	17	0.16	0.07	0.03	0.06	0.05	0.08	0.09	0.07	0.09	0.04	0.04	0.07	0.04	0.09	0.03
Firezza	3.23	16	0.18	0.08	0.03	0.06	0.06	0.07	0.10	0.05	0.09	0.04	0.04	0.07	0.04	0.07	0.03
Big John's	3.26	13	0.18	0.07	0.03	0.06	0.06	0.08	0.09	0.05	0.09	0.03	0.04	0.07	0.04	0.07	0.03
Caspian Pizza	3.60	5	0.16	0.07	0.02	0.06	0.06	0.09	0.10	0.05	0.08	0.04	0.04	0.07	0.04	0.08	0.03
wagamama	3.39	4	0.16	0.07	0.02	0.06	0.06	0.07	0.10	0.06	0.09	0.03	0.04	0.07	0.04	0.08	0.03
Anova F value (df = 30)			**88.82*****	**9.27*****	**21.32*****	**55.15*****	**43.63*****	**74.86*****	**20.15*****	**136.51*****	**64.58*****	**82.61*****	**339.6*****	**31.87*****	**15.57*****	**109.21*****	**110.39*****

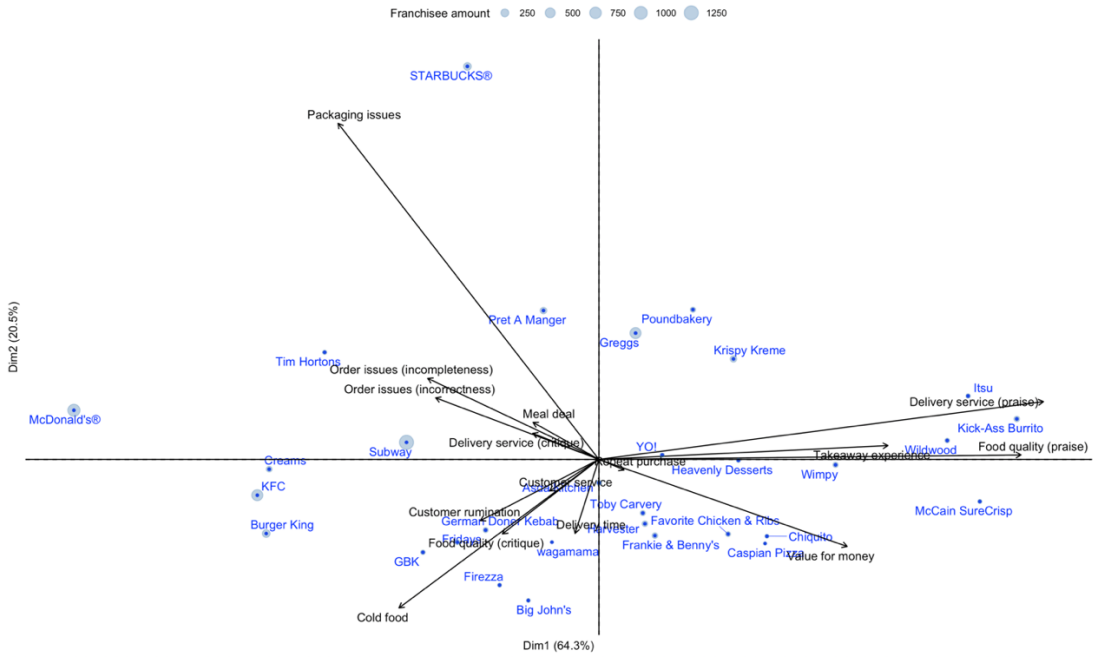

Figure 28 Expected topic proportion distribution across brands

As we can see from the figure, the dots present all brands from franchised restaurant with blue graded circles representing the size of the franchise (the amount of the franchisee on JustEat). There are several relatively large franchising chains such as McDonald's, Subway, and KFC, which are closer to the negative dimensions including the order incompleteness and order issues. Some brands (German Doner Kebab, Wagamama, etc.) are more likely to have issues related to the food quality and complain about the delivery time. In contrast, some brands are more likely to receive the praise from customers regarding delivery service, food quality or the overall takeaway experience, such as Itsu and Wildwood. In particular, 'starbucks' is highly likely to occur with packaging issues. It could be because of the product type which requires higher requirement for the package.

4.5 Discussion and implications

In the OFD sector, both franchised and independent businesses place a premium on customer feedback. From the standpoint of the service supply chain, they yearn for improved management to produce and provide goods and services that meet customer expectations. Our study compiled customer reviews from *JustEat* and extracted the latent determinants of customer satisfaction with the process. The results indicate that

customers are primarily concerned about food quality, delivery performance and issues relating to processing, largely in accordance with the findings from previous study (Xu, 2021). Our key finding though is that reviews of independent restaurants are more likely to include positive topics than reviews of franchised ones. This is in contrast with Barthélemy et al. (2021) in the hotel industry, who found that customer were more satisfied with chain-affiliated units. Customers of franchised restaurants are significantly more likely to complain about order issues and the delivery of a wrong meal, while independent restaurants are more likely to receive praise regarding the quality of their food quality but critiqued regarding delivery time. In addition, independent restaurants seem to have received reviews of stronger valence.

4.5.1 Theoretical implications

This study contributes to theory in several ways. First, with regard to the trichotomy of the actors involved in the franchising strategy (Dant, 2008)– franchisor, franchisee and customer – this work adopted the customer perspective, while past research almost invariable has adopted the franchisor (e.g., ownership structure) and franchisee perspectives (e.g., business format, performance). We thus extend the extant franchising literature, which has almost exclusively evaluated the strategy while focussing at the franchisor or franchisee level (Nijmeijer, Fabbricotti and Huijsman, 2014) and measuring outcomes such as survival, growth and financial performance. In the context of the ODF sector, our work examined customer satisfaction with restaurants, a precursor of firm-level performance outcomes (Symitsi et al., 2021). Using secondary data voluntarily submitted by thousands of customers, we further compared the determinants of customer satisfaction between franchised and independent restaurants.

It is this aspect that gives rise to our second theoretical contribution. Compared to previous studies that examined customer satisfaction towards franchised and independent businesses (Banerjee and Chua, 2016; Moreno-Perdigón, Guzmán-Pérez and Mesa, 2021), our study exhibits the following novelties. First, our data (collected from OFD platforms) might be less biased than data analysed in these past studies (which primarily came from the hotel sector). This is because reviews of hotels are more likely to come from single purchase or low-frequency purchase customers. It is reasonable to assume that online food ordering is a higher-frequency and repeated

purchase, making an evaluation of customer satisfaction with such restaurants more reliable. Second, the dimensions across which we compared the two types of restaurants were latent, and extracted from customers' freely submitted textual comments. This is in contrast with the pre-specified aspects that online hotel booking platforms 'force' customers to rate. Customers might attach asymmetric importance to a given sub-areas of their experience, something that is lost if one solely looks at quantitative ratings of specific dimensions. A topic modelling approach to analysing free text, like the one utilised in this study, can identify the dimensions that customers are most concerned with, and provide a more nuanced comparison of customer satisfaction with independent versus franchised businesses. It can further allow one to account for and understand the influence of unit, brand and customer-level characteristics.

4.5.2 Implications for management practice

These findings provide insights to managers of both franchised and independent businesses. First, it enables them to ascertain the heterogeneity of the extracted dimensions; since restaurants are overwhelmed with a large volume of customer reviews, it is generally difficult for them to identify the key insights without the appropriate infrastructure and expertise. Our method identified the dimensions customers are more concerned with, leading to customer (dis)satisfaction. For instance, critique about delivery time emerges as a strong differentiating factor of reviews with low rating ('1' versus '6'), suggesting that restaurants that aim at improving customer satisfaction should place the utmost importance (e.g., management attention, financial resources) in minimizing delays in order preparation and delivery. Restaurant managers can assess their performance across the identified dimensions and proceed to more informed interventions that will make their service supply chain more effective, and subsequently improve their review ratings. This could also enable them to pursue differentiation-oriented strategies by advertising their specific strengths (e.g., food quality, delivery time) to achieve competitive advantage (Porter and Strategy, 1980).

Moreover, our findings suggest that compared to independent restaurants, customers of franchised restaurants are more likely to receive the wrong meal or have issues with the order (e.g., missing item) and delivery. This is important, since order

process management has a great impact on customers' perception of the service, and is crucial for maintaining customer loyalty (Baltacioglu et al., 2007; Ellram, Tate, and Billington, 2004). Thus, managers of franchise chains might need to consider investment in training for employees of franchise units and optimisation of the ordering process. If those employees are unable to consistently follow the set procedures and abide by the franchisor's standards, the objectives of the franchising strategy – standardisation and efficiency – might be compromised.

4.6 Limitations and future research

Our work has limitations that can be addressed in future research. First, one limitation is the single source of datasets. Our sample is collected from the restaurant industry only while the franchising strategy applied in other service industries such as hotels could behave different. In addition, even JustEat is the main OFD service provider in the UK market, there are some other popular OFD platforms that can be included. Second, the topic solutions could be affected by other review-specific covariates, such as the type of cuisine, which we didn't include in the topic model. For instance, customers might not have the same expectations when ordering a cup of coffee versus ordering a seafood dish. Another limitation is the adoption of STM, from which we interpret the topic solution using human experience. We utilised several diagnostic metrics to find the optimal STM model, however, the bias still exists when we try to label each topic based on the top loading words as the labels are generated by researchers.

The further research could consider the following directions. First, the geographical location of the restaurants should be included in the analysis, especially for the franchised restaurants. It should be considered while collecting the datasets and including is as a covariate in the analysis since the monitoring cost involved in managing geographically dispersed units is an important driver for adopting the franchising strategy. Second, after identifying the customers satisfaction and dissatisfaction from the reviews, future research could consider the reactions from the business regarding these dimensions and explore how they establish the service recovery.

CHAPTER 5 Mapping the signalling environment in sustainability-focused entrepreneurship and linking it to investment inputs: a topic modelling approach on company self-descriptions

5.1 Introduction

The growing emphasis on sustainable technologies suggests an evolutionary shift towards cleaner production venturing (Hegeman and Sørheim, 2021; Shahzad et al., 2022). The level of attention given to support commercialization of sustainability initiatives has mushroomed, despite the risks associated with funding such complex endeavours (Mazzucato and Semieniuk, 2018; Stern and Valero, 2021).With investment in climate-focused technology reaching ¢50B USD across 600+ deals in 2022[3] , twice the size of capital deployed in 2021, the intertwined realities of sustainable energy, high technology and social impact depend upon the capacity of entrepreneurs to develop actionable responses to address the climate challenge (George, Merrill and Schillebeeckx, 2021; Horne and Fichter, 2022). In this context, new ventures assume a seminal role in introducing high-technology solutions to concurrently support decarbonization (Bergset and Fichter, 2015; Freeman, 2019).

However, a successful transition from a new venture to an "*impact start-up*", able to scale sustainability-oriented technologies is dependent on the firm's ability to overcome liabilities of newness and smallness, which suggest a greater risk of failure (Audretsch, Bönte and Mahagaonkar, 2012; Lee et al., 2012; Gimenez-Fernandez, Sandulli and Bogers, 2020). Liabilities of its kind indicate less control of productive resources (Eggers, 2020) and limited access to funding riskier investments (Brown and Lee, 2019). To this end, funding gaps are acknowledged as a key barrier to entrepreneurial growth (Cavallo et al., 2020). In investigating factors that affect funding access, information asymmetry has been identified as a key contaminant in relationships between firms and investors (Bergset, 2015).

As such, investment decisions in sustainability funding in early stage and series funding (Islam, Fremeth and Marcus, 2018; Pollock et al., 2010) may be riddled in inconsistencies related to available information. This is understandable, owing to

[3] https://tinyurl.com/2p9pnwwn

the knowledge problems associated with the entrepreneurial act (Fisher and Neubert, 2022). Asymmetries may be more pronounced in firms that are in the early phases of developing sustainable technologies, where entrepreneurs are developing value propositions that are not yet considered as important to investors or technologies that are not yet understood.

In response to asymmetries, stakeholders search for signals, the latent qualities that provide information about attributes and likely outcomes (Spence, 1978; Bergh et al., 2014). In entrepreneurship literature, there is little understanding of what and how to signal to prospective investors to gain investment attention (Colombo, 2021) and the debate remains in its infancy in sustainability-focused entrepreneurial venturing. This is a critical neglect, as the nature of '*deep-tech'*, disruptive solutions represent a distinctive characteristic of sustainable technologies and assume a long-term development cycle (Harrer and Owen, 2022). To this end, the study suggests that digital enablers (Akbari and Hopkins, 2022; Kimjeon and Davidsson, 2022) may mitigate the impact of asymmetries and provide a research avenue for examining the Signalling environment in sustainability-focused entrepreneurship (Aben et al., 2021). This is important, as the enabling role of digitalization in enhancing circular economy outcomes is gaining momentum in sustainability research (Neligan et al., 2023). As such, the explanatory value of signals depicted from information and knowledge related to firm descriptors and its investment input implications, cannot be underscored.

The present study introduces a novel approach in mapping the Signalling environment in sustainability entrepreneurship. Utilizing a large sample of 2,300 firms over a period of 10 years collected through the CrunchBase platform, topic modeling procedures were employed to unmask latent signals presented in company descriptions and areas of sustainability activities. In turn, these were then classified via task and institutional elements and their impact on investment inputs was evaluated. The study makes significant additions in two interrelated areas of intellectual inquiry. First, our study contributes to Signalling theory by presenting the complex dynamics of the Signalling environment in sustainability-focused entrepreneurship. Sustainable start-ups profess their qualities not only through firm-level practices but also through institutional and task-related signals. Second, we examine the impact of these signals on investment inputs. The study's findings

suggest that certain signals have positive influence upon investment decisions, which leads to increased investment inputs.

The rest of the paper is organized as follows. The next section presents the study's theoretical framework, where we discuss the nature of information asymmetries in the context of sustainability-oriented venturing and the Signalling framework to exemplify the task and institutional signals relevant to sustainability-focused technology ventures. This is followed by a discussion of the dataset and the study's analytical treatment. We then discuss the study's key findings and present the study's implications for sustainability technology firms, investors and policy-makers.

5.2 Theoretical framework

5.2.1 Sustainable technologies and information asymmetries

Sustainability-oriented entrepreneurship envisages the pursuit of opportunities with positive environmental tradeoffs. Sustainable technologies, typically brought to market by start-ups and small firms, hold the promise of disrupting existing industries and in addition to implications for socially responsible venturing, may also give rise to substantial payoffs to investors (Mimno et al., 2011; Kwon, Lim and Lee, 2018). To demonstrate their value to external entities, firms must encompass a venturing context riddled in information asymmetries, complexities, and unforeseeable uncertainty. Sustainable technology projects are fundamentally ambiguous and bear considerable risks in terms of intra-industry competitive dynamics, pace of disruption (Hart and Milstein, 1999), shortened life cycle of products, and challenges in consumer adoption (Cumming, Leboeuf and Schwienbacher, 2017; Li, Wu and Mai, 2019).

In such a context, technology assessment is necessary (Gladysz and Kluczek, 2017), yet reliant on novel and complex sources of asymmetric information for investors. This is a perennial challenge, considering key skepticisms about the cleaner production impact of proposed projects *and* difficulties to identify where market opportunities for sustainability-focused investments currently reside (Apostolopoulos et al., 2020). The main outcomes of asymmetrical information have been principally discussed in entrepreneurship literature with regards to agency costs (e.g. Norton, 1996), financial assessment of risk and future profitability (Bergset and Fichter, 2015),

firm capital constrains (Mrkajic, Murtinu and Scalera, 2019) and the role of Governmental support in choosing market winners (Pitelis, Vasilakos and Chalvatzis, 2020). As such, it is important to examine the nature of asymmetries in the sustainable technologies debate to understand whether sustainability funding input is driven by an underlying dynamic with regards to the firm's signalling environment.

Extant literature discusses the characteristics of information asymmetry by examining the impact of signals under different levels of information access in discursive market conditions (Hsu and Ziedonis, 2013; Reuer, Tong and Wu, 2012; Courtney, Dutta and Li, 2017). Asymmetries of its kind can be traced at the equivocality stage (Dahlmann and Roehrich, 2019; Daft and Lengel, 1986), where information ambiguity is prevalent, leading to conflicting priorities between evaluating project risk and interpreting investment potential. Equivocality describes the "*…existence of multiple meanings or interpretations in a specific context, each of which is individually unambiguous, but collectively, they are either mutually exclusive or in conflict with one another*" (Fisher and Neubert, 2022:p.4).

Asymmetries in the sustainability context suggest misalignment of conflicting and unstructured sources of data that make information processing crucial (Harrer and Owen, 2022; Jolink and Niesten, 2021; Dahlmann and Roehrich, 2019). By implication, this suggests that funding providers must make resource provision decisions based on limited information related to the viability of such an entrepreneurial endeavor. To better realize equivocality boundaries in this context, Signalling theory offers an explanatory lens to determine how organizations convey information about themselves to external parties (Truong et al., 2022) and a foundation to interpret signals that reduce uncertainty associated with making selections among a choice in situations with inconsistently distributed information (Spence, 1978)

5.2.2 Signalling actions and sustainability-focused entrepreneurship

Signals can be broadly classified in line with their characteristics, costs, and intentional actions (Bafera and Kleinert, 2022). Effectiveness of Signalling actions is moderated by the industry of the new venture (Hsu, 2007) and Signalling strength respectively (Tumasjan, Braun and Stolz, 2021), whereas investors interpret the same signals in a different manner. In depicting the importance of Signalling actions,

entrepreneurship research suggests a key distinction among economic and quality signals with the latter debate demonstrating a respective focus on productive quality signals. In addition, rhetorical signals suggest a recent and timely addition in the Signalling debate, professing the relevance of language-based information that helps potential investors determine funding decisions (Steigenberger and Wilhelm, 2018).

In explaining selection decisions, past work looked at economic signals that communicate the ability to generate financial returns as interpreted from business angels (Prasad, 2000), venture capitalists (Plummer, Allison and Connelly, 2016) and financial institutions (Eddleston et al., 2016). On the other hand, research examined the unobservable potential for funding success in terms of new venture quality signals such as human capital (Wiklund and Shepherd, 2003; Baum and Silverman, 2004; Beckman, Burton and O'Reilly, 2007), intellectual capital (Block et al., 2014) and firm age and size (e.g. Baum et al., 2000). In terms of productive quality signals, research stresses the importance of the firm's collaborative capacity. In this instance, alliances and collaborations represent a productive capacity signal (Hoenig and Henkel, 2015) demonstrating the firm's capitalization of its networking capability. Alliance capital allows sustainable start-ups to gain access to complementary resources (Chung, Singh and Lee, 2000) and represent an observable quality of the firm's technological legitimacy (Baum and Silverman, 2004). Collaborations, demonstrate the firm's potential for market acceptance and financial returns (Block et al., 2014; Gans, Hsu and Stern, 2000). As an asset, it enables a sustainability start-up to develop know-how and expertise. Masini and Menichetti (2012) provided important evidence to suggest that investor's a-priori beliefs, preferences over policy instruments and attitude towards technological risk affect the likelihood of investing in such projects.

Research suggests that investors are keen to support start-ups that signal fast and secure returns than long-term technology development (Masini and Menichetti, 2012). For example, Hoenig and Henkel (2015) identified patents, alliances and team experience as key institutional signals for venture capital funding. Patents and trademarks represent indicators of esteem in terms of innovation activities (Mendonca et al., 2004). Studies suggest that patents act as a source of credibility underpinning the value of the innovation (Hsu and Ziedonis, 2013; Audretsch, Bönte and Mahagaonkar, 2012), but with varying impact in early rounds of capital financing (Hoenen et al., 2014). Trademarks, referring to "*...a word, phrase, symbol, or logo",*

represent a key point of differentiation among a firm and its competitors (Block et al., 2014:527) with research suggesting that trademarks represent the marketing side of an innovation. However, in the sustainability context, such signals may have paradoxical or even diminishing effects in investment inputs, when considering the long-term outcome of proposed technologies (Harrer and Owen, 2022) and asymmetries of this kind may drive underfunding in sustainability-focused start-ups (Gaddy et al., 2017).

Building on the above sentiments, the Signalling context of sustainability-focused firms can be examined in two interrelated areas of activity, the task and institutional environment (Connelly et al., 2011). The task environment refers to activities undertaken by sustainability-focused technology start-ups, whereas institutional signals depict organizational characteristics that firms utilize as indicators of esteem. As such, the present study disentangles the signals that are most prevalent in sustainability-focused start-ups and the following sections further elaborate task and institutional signals on the merit of their characteristics, costs and intentionality (Bafera and Kleinert, 2022) leading to theorization on the value of the joint effects among task and institutional signals.

5.2.3 Task signals

Task signals assume the central characteristic of observability (Connelly et al., 2011), which suggests the extent to which the signal is noticed and understood by potential investors (Ahlers et al., 2015). Such signals reflect activities related to projects, indicating higher venture quality. In early stages of development, observability is associated with the proof points that demonstrate the firm's commitment to commercialize technologies with uncertain merits (Islam, Fremeth and Marcus, 2018). In a new firm context, where there is no proven track record of successful project development and technological novelty is commonplace, the capacity to convey the 'right' signals irrespective of equivocality limitations may provide additional explanatory value to stakeholders for making sound investment decisions (Giones and Francesc, 2015). This is consistent with the value of deliberate experimentation as purposeful interaction, for achieving strategic legitimacy (Bojovic, Genet and Sabatier, 2018). In the empirical context of this investigation, firm and activity signals reflect critical information (Steigenberger and Wilhelm, 2018), reflecting the firm's fit and business interests in innovative activities related to an emerging sector context

(Criscuolo, Nicolaou and Salter, 2012; Kimjeon and Davidsson, 2022). The proposed innovations fall under the sustainable technology agenda such as renewable energy, recycling, green transportation, buildings, electric vehicles, chemistry, lighting (Marra, Antonelli and Pozzi, 2017). Commitment to sustainable activities signals engagement in costly processes that deliver higher value using fewer resources and producing less pollution than current standards (Cooke, 2008). Research notes that internal task signals can be observed on a broad range of product, market, team and financial characteristics (Bafera and Kleinert, 2022).

Product characteristics relate to protected products or services that the firm currently offers that align with sustainability goals. This is important as in the broader context of technology ventures, product characteristics compliment the value of market or investment signals (Bapna, 2019). Market characteristics reflect market opportunities related to carbon emission reduction or promotion of circular economy. Team characteristics point towards entrepreneurial skills and expertise relevant to the renewable economy, while financial characteristics refer to current and future sources of funding. To this end, task signals that reflect sustainability actions manifest signal intentionality (Vanacker and Forbes, 2016), by strategically positioning the firm in emerging sectors, allowing it to transition towards the latter stages of development.

5.2.4 Institutional signals

In understanding venture quality, institutional-level signals represent a quality indicator due to costs associated with development of novel activities (Connelly et al., 2011). These endeavors reflect signal-worthy options, demonstrating the strategic use of indicators that demonstrate the allocation of valuable resources (Islam et al., 2018). This is an important depiction of the institutional Signalling logic that characterizes sustainability venturing, when considering the nature of their developmental activities that is preoccupied with technology development and strategic positioning as options that bear long-term Signalling value. To this end, Signalling costs can be observed on rhetorical signals that demonstrate how start-ups describe their vision, advances, and operational model (e.g. Steigenberger and Wilhelm, 2018; Payne et al., 2013). Innovative start-ups must overcome difficulties of new organizations (Stinchcombe, 1965), in terms of formulating routines and are also characterized by unstable links to customers, suppliers and partners (Gimenez-Fernandez, Sandulli and Bogers, 2020). From a Signalling perspective, newness may represent an institutional-level

characteristic that is advantageous in terms of business model flexibility(Kapoor and Klueter, 2015) and presents fewer limitations based on lower risk aversion (Audretsch and Keilbach, 2007). In addition to newness, start-ups also must deal with lack of resources necessary for business development and difficult access to resources necessary to grow.

Research suggests that institutional characteristics may signal the firm's underlying qualities (Courtney et al., 2017), with investors seeking evidence related to the future success of the business, through organizational signals (Pinelli, Davachi and Higgins, 2022). In examining the unobservable potential for funding success, new venture quality signals that portray the firm's sustainable considerations (Connelly et al., 2011) represent a possible differentiator. In addition, Loock (2012) stressed investor preference for service-driven business models for renewable energy, addressing customer needs rather than technology or price. One way to mitigate asymmetries through institutional signals, is to convey information on the firm's qualities and intentions in the form of credible signals (Bafera and Kleinert, 2022) and there is recent work following that line of reasoning. Cowan and Guzman (2020) suggest that sustainability signals increase corporate brand performance and brand equity in domestic environments. Bento, Gianfrate and Thoni (2019) identified that perceived sustainable mission positively influences crowdfunding outcomes in sustainability-oriented campaigns. In a similar context, Wehnert and Beckmann (2021) depicted the centrality of sustainability orientation in mitigating information asymmetries.

5.2.5 Task and Institutional signal consistency

By Signalling, firms portray information on their qualities and intentions to support investment decision-making (Elitzur and Gavious, 2003). To overcome concerns about information equivocality and mitigate asymmetries, firms may also increase signal consistency, which refers to the joint effect of multiple signals from one sender (Colombo, 2021). This conceptual depiction is of particular importance in sustainability ventures. Transmission of multiple signals may have a positive, additive effect when signals demonstrate quality in different domains and avoid replication (Colombo, Meoli and Vismara, 2019). The transmission of several observable signals

(Connelly et al., 2011), may also provide firms with marginal benefits regarding investment inputs.

Research exploring joint Signalling effects is paradigmatically limited, yet suggestive of a complex relationship on how multiple signals add value (Pollock et al., 2010). In the context of sustainability-focused entrepreneurship, the interplay between task and institutional signals consistently communicates firm qualities and characteristics and may very well sharpen the understanding of investors on where future opportunities for investment can be identified. This is important, as potential investors are presented with the opportunity to evaluate firm potential by overcoming the boundary conditions of multiple Signalling that leads to signal conflict (Chan et al., 2020) and determine venture quality through complementary sources of information (Courtney, Dutta and Li, 2017; Chen et al., 2018). In addition, considering the rather noisy industry environment and value proposition ambiguity associated with sustainability initiatives, the interplay between different signals may also present an amplification effect where one signal increases attention to other signals (Bafera and Kleinert, 2022; Steigenberger and Wilhelm, 2018; Vanacker and Forbes, 2016).

5.2.6 Investment inputs

Research depicted the links between a firm's Signalling environment and resource acquisition (Connelly et al., 2011) and notable work further expanded on this assertion to explore the essential connection between signals and investment inputs (Islam et al., 2018). The sustainable finance investment domain is characterized from fragmentation and lack of consensus with respect to its outcomes (Cunha, Meira and Orsato, 2021). Sustainability-focused investing is characterized from high investment risks (Hegeman and Sørheim, 2021) substantial initial capital requirements and commercializing with long lead times (Islam et al., 2018). The typical early investment model is consistent with technology development (Bürer and Wüstenhagen, 2009) and can be broadly classified on the basis of identifying investment inputs at *seed* and *series* funding. Seed funding is related to early-stage investment, usually linked to prototype development, where the vast majority of funding comes through informal sources such as business angels (Barringer and Ireland, 2010). Once a market dynamic is accomplished, then venture capital and

private equity investment is attracted (Puri and Zarutskie, 2012), leading to series funding which is linked to venture expansion (Islam, Fremeth and Marcus, 2018) and venture capitalist involvement respectively.

The opportunity to attract initial funding from venture capitalists contributes as a positive signal demonstrating credibility towards commercialization (Islam et al., 2018). However, the vast majority of previous work examined the effects of Signalling at the Initial Public Offering (IPO) and post-IPO stages (Payne et al., 2013; Chen et al., 2018; Colombo et al., 2019; Wang et al., 2019), with a similar pattern observed in sustainability-related research (Harasheh, 2022a, 2022b; Kang and Lam, 2023). This is inherently problematic as in the case of sustainability-focused ventures, research is paradigmatically silent in understanding resource acquisition in early stages of venture formation (Ko and McKelvie, 2018). This is a critical neglect considering that theory suggests that the strength of the relationship between Signalling and investment outcomes seems to be higher in conditions of uncertainty, when asymmetric information between firms and investors are also high (Ko and McKelvie, 2018).

5.2.7 Research gaps

After reviewing the literature, there are several research gaps that can be filled by this study. First, the existing literature acknowledges the existence of information asymmetries in sustainable technology ventures. There is a lack of knowledge regarding the specific impact of these asymmetries on the decision-making processes and how signal works at various stages of investments. The review indicates that investors interpret diverse signals (economic, qualitative, rhetorical) in varying ways, affecting their efficacy and perception. However, there is a dearth of comprehensive empirical data about the efficacy of these signals in mitigating information asymmetries, particularly within the field of sustainability. This study could provide the quantification of the influence that various forms of signals have on investment results.

Besides, existing literature highlights the significance of signalling in the sustainability sector but lacks in-depth analysis of the potential variations in signalling dynamics across all task and institutional signals. The thesis could examine the interplay between task and institutional signals, highlighting the potential for

investigating the interaction and combined impact of these signals on investment choices. This entails evaluating the internal relationships across these signals.

What's more, the literature discusses the signalling impacts but fails to provide a longitudinal analysis of how the value and perception of signals change as a venture progresses from a startup to a more established corporation. Subsequent research could explore the effects of initial signals and the changing characteristics of signals as the company expands. Early-stage sustainable initiatives face a significant challenge in effectively signalling their potential to attract initial investments, especially before reaching the stages that often attract more formal venture capital or equity funding. Knowing how these ventures manage signalling especially in the initial phases can be crucial for startups.

5.3 Data and methods

5.3.1 Dataset description

Data were sourced from CrunchBase (Ferrati and Muffatto, 2020; Felgueiras, Batista and Carvalho, 2020). As a database for acquiring and categorizing business information including investments, funding information, acquisitions, regarding private and public companies in various areas of economic activity, the platform is notably used in sustainability-focused entrepreneurship research (Marra, Carlei and Baldassari, 2020; Marra, Antonelli and Pozzi, 2017). In our case, the use of the CrunchBase platform enables all relevant agents to signal strategically actionable knowledge which in turn, impacts information equivocality and may facilitate interaction and knowledge exchange. CrunchBase obtains the data in four ways: the Crunchbase community, an in-house data team, AI and machine learning, and the Crunchbase venture program (Dalle, Besten and Menon, 2017). Companies founded from 2005 to 2022 were selected. We collected the information of 5,099 companies from the sustainable technologies industry group encompassing 19 specific industries organized by Crunchbase.

Table 25 Distribution of companies across 19 industries within the sustainable technologies group

Industry	Number of Companies	% of Total
Renewable Energy	1775	34.81%
Solar	776	15.22%
Sustainability	759	14.89%
CleanTech	740	14.51%
Energy Efficiency	580	11.37%
Clean Energy	520	10.20%
Waste Management	400	7.84%
Environmental Engineering	389	7.63%
GreenTech	341	6.69%
Recycling	335	6.57%
Natural Resources	269	5.28%
Organic	209	4.10%
Water Purification	199	3.90%
Wind Energy	106	2.08%
Green Consumer Goods	76	1.49%
Green Building	71	1.39%
Biofuel	54	1.06%
Biomass Energy	51	1.00%
Pollution Control	42	0.82%

There are 19 specific sustainability-related industries categorised by CrunchBase. Table 25 displays the distribution of sustainable technologies companies across 19 industries within the sustainability industry group. The most companies are from the renewable energy industry, amounting to 1,775 (34.81% of all), followed by Solar (776) and Sustainability (759). The least number of companies occurs in Pollution Control industry (42), Biomass Energy industry (51), and Biofuel industry (54).

The dataset contained organizations' names, their full descriptions, specific industry, operational elements (number of employees, year of incorporation) and investment inputs (founding rounds and total funding amount). Table 26 summarises the distributions of operational elements and text length in our sample. Number of employees were counted as a discrete ordinal variable with companies having 1-10 employees accounting for the majority of the dataset (44.04% of the total). More

mature companies with more than 11 employees (11-50) represent the other significant part of the sample (39.78%), suggesting a skewness to small and micro companies (EU, 2003). The rest of the sample was comprised of companies having 51-100 (6.45%), 101-250 (5.35%), 251-500 (2.14%), 501-1000 (1.20%), 1001-5000 (0.63%), 5001-10000 (0.18%), and more than 10001 employees (0.24%).

Table 26 Summary statistics of the sample

Variables	**Mean**	**Median**	**SD**	**Min**	**Max**
Years of incorporation	9.43	9	4.32	1	18
Funding amount (USD)	43,720,613	2,434,426	298,470,802	1000	15,811,577,400
Funding rounds	2.72	2	2.45	1	34
Length of company description (words)	70.64	58	55.80	3	697
		Median	**% of Total**	**Min (%)**	**Max (%)**
Number of Employees (Band)		11-50	39.78%	1-10 (44.04%)	10000+1 (0.24%)

5.3.2 Keyword extraction using Textrank

Keyword extraction refers to the systematic identification of key terms or key phrases from a document that are chosen to accurately represent the subject matter of the document (Beliga, Mestrovic, and Martincic-Ipsic, 2015). As the inexhaustible source of information, the textual data on the Web is continually expanding. This expansion results in a growing number of digital documents, making manual keyword extraction impractical. Thus, keyword extraction emerges as a fundamental endeavour in the fields of text mining, natural language processing, and information retrieval (Onan, 2016). The condensed representation provided by keyword extraction has implication for a variety of applications, including automated indexing, text summarisation and clustering, which all benefit from this process (Zhang et al., 2008).

Automated keyword generation incorporates two subcategories: keyword assignment and keyword extraction (Siddiqi and Sharan, 2015). In the task of keyword assignment, a group of prospective keywords is selected from a controlled lexicon while keyword extraction identifies the most relevant terms within the subject document (Beliga et al., 2015). Compared with the keyword assignment, keyword extraction is considered in this study as it is less costly, time saving to extract the representative terms from the corpus. Approaches to keyword extraction could be achieved using several approaches, such as the statistical approaches, linguistic approaches, machine learning models and some other approaches (Han and Kamber, 2006).

The objective of a keyword extraction application is to automatically identify a set of terms within a text that accurately summarise the content of a document. These identified keywords have the potential to be useful in multiple contexts such as serving as a concise representation of a specific document, which is the concept in text summarisation.

TextRank algorithm is introduced by (Mihalcea and Tarau) 2004 as a graph-based ranking model optimised for text processing. This algorithm is applicable to numerous natural language processing tasks, such as the keywords extraction and sentences extraction from text documents. The Textrank ranking model will be deployed on the graph structure. Graph-based ranking approaches concentrate on assessing the importance of nodes or vertices within a graph, using information recursively garnered from the entire graph. This determination is dependent on the integration of global information that is derived iteratively from the entire graph. A graph-based ranking model's underlying principle can be compared to "voting" or "recommendation."

In this context, when a node or vertex forms a connection with another, it is essentially endorsing the latter one. The weight of this endorsement corresponds to the number of votes received by a node. Thus, an increase in the number of endorsements increases the prominence of the node. The ranking model attentively considers this complex relationship between node importance and endorsing influence. Therefore, the evaluation of a node's score depends on both the number of endorsements it has received and the collective influence of its endorsers.

Textrank model could be represented as a directed graph $G = (V, E)$, which consists of a set of nodes (V) and set of edges E, the edges E is the subset of $V * V$.

for a given node V_i, $In(V_i)$ represents the set of nodes that point to the node V_i, while $Out(V_i)$ indicates the set of nodes that the node $V_i$points to other nodes. The following formula is to determine the score of the node V_i, shown as below:

$$S(V_i) = (1-d) + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j) \quad (43)$$

where d could be considered as a damping factor that are set between 0 and 1. The initial value is required to specify in order to calculate the score of the nodes of the graph. Then the score will be computed recursively until convergence.

When TextRank is used for keyword extraction, the method involves dividing the text into tokens and assigning POS tags to each token. Subsequently, a node is created in the graph for each lexical unit satisfying the syntactic filter, and edges are created between nodes when words co-occur within a specified window of words. This procedure generates an undirected and unweighted representation of the text's graph.

It is expected to generate a collection of word or phrases using Textrank that encapsulate the given text. Therefore, the sequence of one or more lexical units represented as the nodes derived from the text will be ranked in the text graph. The relations between two lexical units will represent a potentially valuable connection (called an "edge") that can be established between the corresponding nodes. The co-occurrence relation determined by the proximity of word appearances is selected in this model. More precisely, two vertices are connected when the corresponding lexical units appear within a defined interval containing no more than N words. This N value may be chosen from a range of 2 to 10 words. As demonstrated by Mihalcea and Strapparava (2009), co-occurrence links serve as indicators of syntactic cohesion, mirroring the significance assigned to semantic links in tasks such as word sense disambiguation. A syntax-based filtering mechanism can be applied to the nodes incorporated into the graph, incorporating only lexical units of particular portions of speech. We limit the word selections to nouns and adjective.

5.3.3 Corpus pre-processing

To prepare the corpus, the following steps are applied to the textual descriptions of companies: i) tokenization (sentences separated into a list of tokens with word as a

unit), ii) stopwords removal according to the SMART list, iii) after POS (part-of-speech) and lemmatization, choosing nouns, adjectives and adverbs as they state the company's vision and status. By excluding words with low frequency (less than 1% of the total amount of company descriptions), the total amount decreases to 5,088.

5.3.4 Latent dirichlet allocation and Structural topic model

Structural topic modeling was employed to discover the latent signals within companies' descriptions. This approach has considerable computational advantages in extracting latent variables from complex datasets, including several types of algorithms which can expose, discover, and extract the thematic structure from a collection of documents (Vayansky and Kumar, 2020). As a type of Bag-of-Word model, the topic modelling approach disregards the order of words in a document but considers the frequency of occurrence of each term as a significant factor (Blei et al., 2003). In text analysis, topic modelling is utilized to identify the identify the events of themes, which are usually called 'topics'. In topic models, *'topic'* is defined as a distribution over a vocabulary of words representing a theme which could be interpreted semantically (Roberts et al., 2016). It is considered as an unsupervised approach as it could infer the topics instead assume the content of topics under supervision.

There are two varieties of topic models: single-membership models and mixed-membership models. Between two types of topic models, mixed-membership allows each document to cover multiple topics rather than restricted to only one topic in single-membership topic models. The most widely applied mix-membership topic model is Latent Dirichlet Allocation (LDA), proposed by Blei et al. (2003). LDA has been applied by researchers from various fields. By introducing the metadata into LDA, Roberts et al. (2014), proposed structural topic models, which emphasized the influence of the covariates from metadata.

In this study, we employed STM rather than LDA for several reasons. First, what differentiates STM from LDA is the topical prevalence parameters and topical content parameters. In LDA, the parameters that control the how document is associated with topics are shared prior Dirichlet parameters across the corpus. However, the topical prevalence (content) parameter in the Dirichlet distribution is replaced with means parameterized by a linear function of observed covariates

(Roberts et al., 2014), which means the document-level information could be incorporated into the model and influence the document-topic distribution and topic-word distribution. With the help of the incorporation of document-level covariates, STM is theoretically believed to produce more reasonable output (Robert et al., 2016). More importantly, we are able to easily determine how document-topic proportion are varied across different level of covariates.

Second, STM can address another shortage of LDA, which is the lack of examination of the topics interrelation as it can be considered as an extension of the correlated topic model in which the correlations among topics could be examined (Blei and Lafferty, 2007). It could help us to identify the relationships among all topics, thus, to discover how the latent signals interact with each other.

Let's assume a corpus of D company descriptions with each description d indexed as $d \in (1, \dots, D)$. Each description contains w observed words indexed as $n \in (1, \dots, N_d)$, which are from a vocabulary of words, indexed by $v \in \{1, \dots, V\}$. The number of topics K is indexed by $k \in \{1, \dots, K\}$, which needs to be specified as an input before the model estimation process. We display the graphical illustration of STM in Figure 29, the model consists of three components conceptually. The topic prevalence model controls the proportion of each topic contributing to a document. The topical content model allows metadata affecting the word frequency within each topic. The core language model generates actual words for each document using the combination of variation from the topic prevalence model and the topical content model.

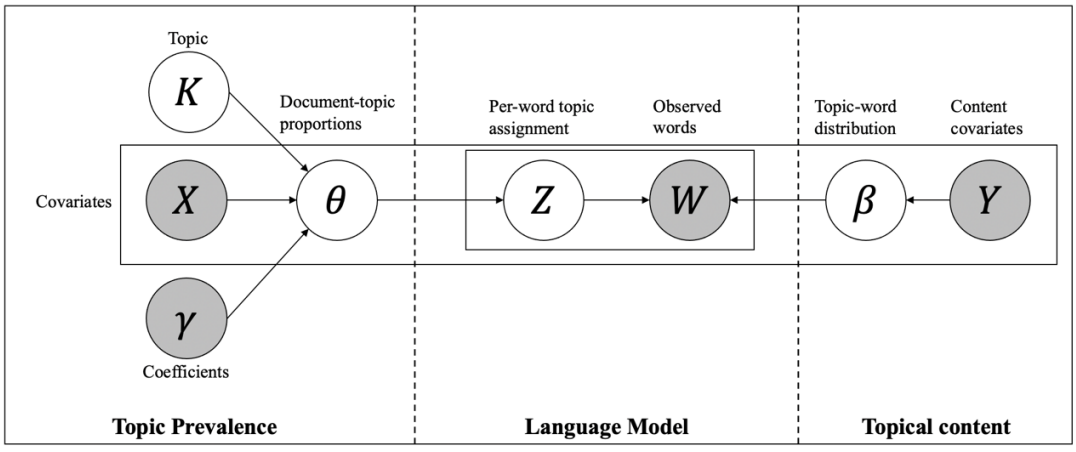

Figure 29 A graphical demonstration of STM (adopted by Roberts et al. 2016)

The text generative process is as follows:

First, the document-level relation to each topic is drawn from a logistic-normal generalized liner model based on covariates X_d.

$$\underset{\theta_d}{\rightarrow} | X_{d\gamma}, \Sigma \sim LogisticNormal(\mu = X_{d\gamma}, \Sigma) \tag{44}$$

where Σ is a $(K-1) \times (K-1)$ covariance matrix. γ represents a $p \times (K-1)$ matrix of coefficients drawn from a normal distribution for each k ($k = 1, \ldots, K-1$) shown as below:

$$\gamma_k \sim \text{Normal}(0, \sigma_k^2) \tag{45}$$

$$\sigma_k^2 \sim \text{InverseGamma(a, b)} \tag{46}$$

where a and b are fixed hyperparameters (Roberts et al., 2016).

Form the description specific distribution over words representing each topic (k) using the baseline word frequency (m), the topic specific derivation $\kappa_k^{(t)}$, as well as the covariate group derivation $\kappa_{y_d}^{(c)}$ and the interaction $\kappa_{y_d,k}^{(i)}$ between them.

$$\beta_{d,k} \propto \exp\left(m + \kappa_k^{(t)} + \kappa_{y_d}^{(c)} + \kappa_{y_d,k}^{(i)}\right) \tag{47}$$

When no topical content covariate is present, β can be simplified as $\beta_{d,k} \propto \exp\left(m + \kappa_k^{(t)}\right)$

For each word $n \in (1, \ldots, N_d)$ in a company description, based on the description-specific distribution over words, drawn the word-specific topic assignment $Z_{d,n}$

$$Z_{d,n} | \underset{\theta_d}{\rightarrow} \sim Mutinomial(\underset{\theta_d}{\rightarrow}) \tag{48}$$

Draw an observed word from the topic chosen.

$$w_{d,n} | z_{d,n}, \beta_{d,k=z_{d,n}} \sim Mutinomial(\beta_{d,k=z_{d,n}}) \tag{49}$$

As shown in the process, STM allows covariates to influence the topic prevalence and topical content, which is the biggest advantage compared to LDA. Instead of prior distribution sharing a global mean, the distribution that controls the document-topic proportions is a logistic Normal distribution and the mean is parameterized as a linear function of covariates. The inclusion of covariates makes the estimation of quantities of interest more accurately and more useful for inference than LDA (Korfiatis et al., 2019; Roberts et al., 2014). Plus, the relationships between the latent topics and covariates could be estimated to explore the marginal effects of

covariates to the topic prevalence. In this study, there are several variables which are employed as covariates to influence the topic prevalence, including year of corporation, funding amount, founding rounds and the number employees.

5.3.5 Estimation of the topic solution

We performed models in R using 'stm' package (Roberts et al., 2019). As previously discussed, STM allows us to introduce the document-level covariates to influence how latent topics distribute. In this study, we included the characteristics of companies (*year of incorporation, funding amount, funding round, and number of employees*) to influence how often a topic is discussed, which is the topic prevalence. To find the optimal number of topics (K), we established and performed models with different K number from 11 to 20. We adopt two metrics to evaluate their performances: held-out likelihood and exclusivity. As shown in the shaded region, we select K number as 17 the optimal for topic number.

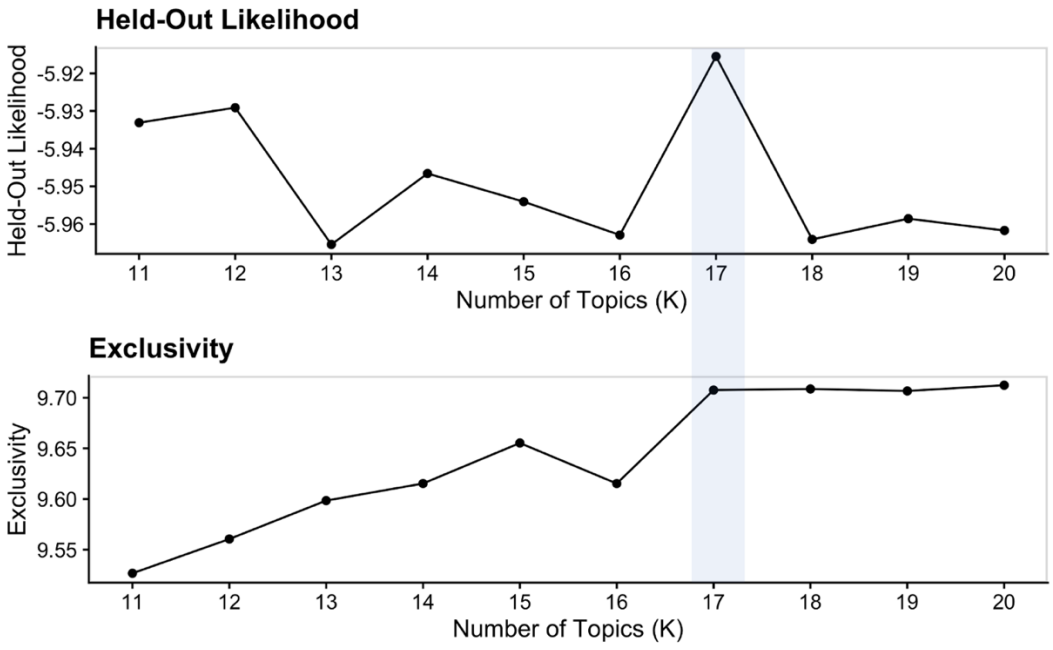

Figure 30 Diagnostic values for topic solution with 11-20 topics

In addition, we adopt the FREX score (Roberts et al., 2016) to measure the topic quality using a combination of frequency and exclusivity.

$$FREX_{k,v} = \left(\frac{\omega}{ECDF\left(\frac{\beta_{k,v}}{\sum_{j=1}^{K}\beta_{j,v}}\right)} + \frac{1-\omega}{ECDF(\beta_{k,v})}\right)^{-1} \quad (50)$$

where ECDF is the empirical cumulative distribution function and we set the weight of exclusivity ω to 0.7 (Korfiatis et al., 2019).

5.4 Results

5.4.1 Text statistics

The histogram plot provides a graphical representation of the distribution of word frequencies obtained from a collection of company descriptions. The x-axis of the histogram represents the frequency count of each word, while the y-axis shows the individual words (their original form after text pre-processing) taken from the corpus.

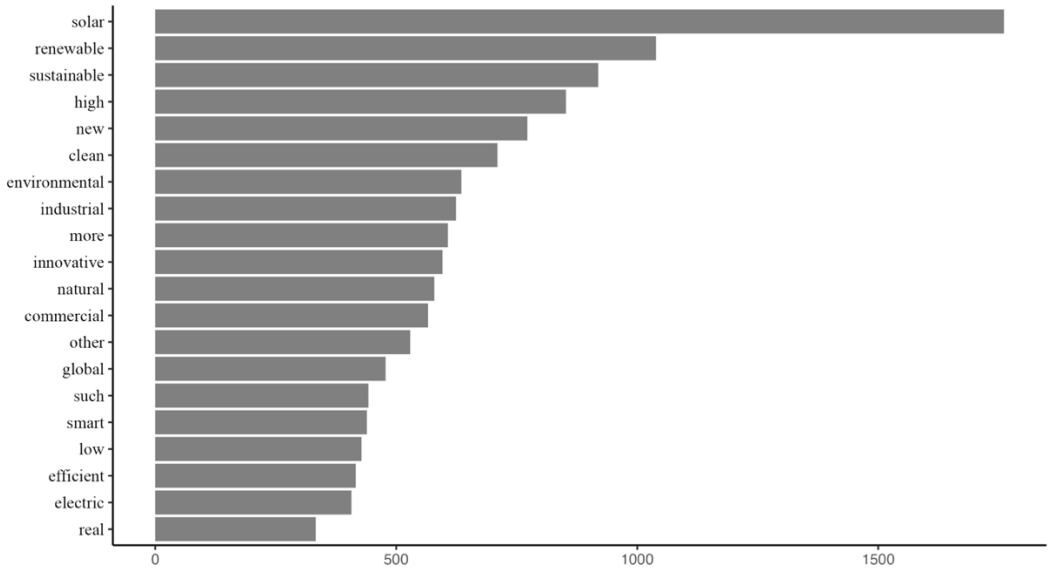

Figure 31 Word histogram for adjectives

It could be shown in Figure 31 that the term 'energy' is the most frequently occurring noun, with a count of 4515. This observation highlights the significant emphasis on energy-related talks in the sustainability sector. The term 'business' is observed with a frequency of 4328, suggesting a recurring mention on organisational entities within the texts. The term 'technology' is seen with a frequency of 2985, indicating the prominence of technical developments in the sustainability discourse of corporate descriptions. Various significant terms, like 'product', 'system', and 'power', exhibit considerable frequencies, thereby emphasising the extensive array of subjects and concepts. These issues span from the provision of products to the

development of system-based solutions and the generation of power.

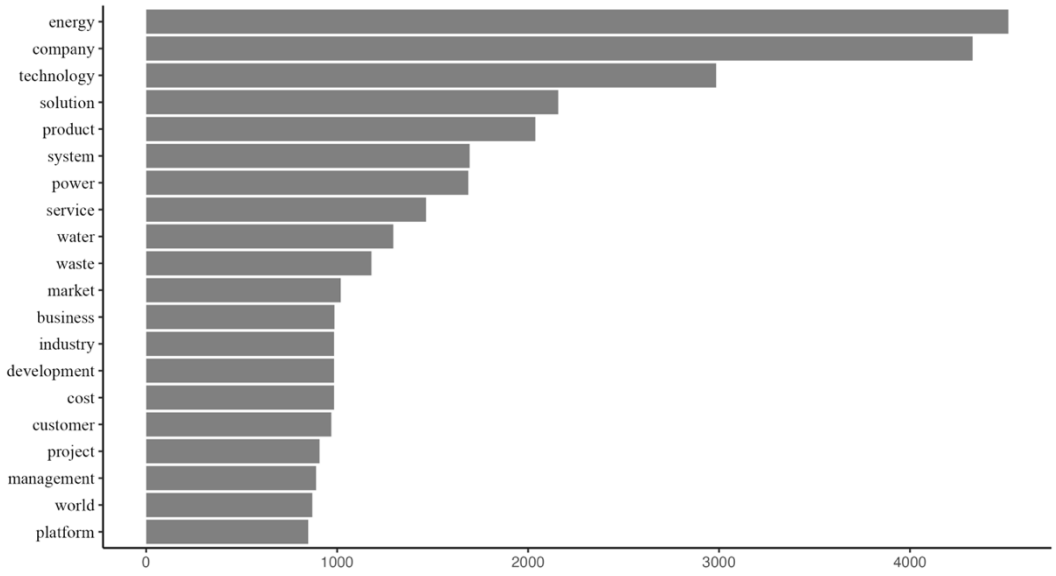

Figure 32 Word histogram for nouns

The frequencies of the terms "water" and "waste," which occur 1294 and 1180 times respectively, serve as examples that highlight the significance placed on environmental resources and waste management. These figures are indicative of the heightened awareness and emphasis on environmental issues within the overarching sustainability framework adopted by organisations. Furthermore, the terms 'market', 'business', 'development', and 'industry', with frequencies ranging from 1019 to 984, indicate the significance placed on market dynamics, business models, initiatives for development, and industrial environments.

Similarly, Figure 32 represents the visual depiction of the frequency distribution of adjectives derived from the corpus. The most frequent adjectives and their corresponding frequencies are shown in the plot. The word ‘solar' holds a predominant position, occurring with a frequency of 1761. This observation highlights a noteworthy emphasis on conversations pertaining to solar-related topics within the sustainability. The term 'renewable' is of significant importance, as seen by its frequency of 1039, which indicates the recurring emphasis on renewable resources and technology within the discourse on sustainability. The term "sustainable" is the third most often used word, appearing 919 times, which highlights the dominant topic of sustainability included in the company descriptions. The frequency of adjectives such as 'environmental', 'industrial', and 'innovation' aims to emphasise the wide array of sustainability factors that are taken into account. These factors might encompass

several aspects, including environmental issues, industrial uses, and inventive methods and solutions.

In summary, the two plots offer a thorough and concise analysis of the most frequently occurring nouns and adjective found in company descriptions. Even it can provide several insights from both the adjectives and nouns, the single word could not demonstrate the clear concepts within the company description. Therefore, we then move to the 2-gram phrase and 3-word phrase using Textrank algorithm.

5.4.2 2-gram and 3-gram keywords extraction

The below 2 tables provides an overview of the key phrases obtained from the company description using the TextRank algorithm, which focused on company descriptions within the domain of sustainability. The TextRank algorithm is a graph-based ranking mechanism commonly employed in text processing to extract significant phrases or concepts from a given text corpus. The table has been structured with three columns. The column "keyword" displays the two-word key phrases that have been extracted. These phrases represent the frequent concepts or phrases that have been mentioned by the company from sustainability industry. The column 'ngram' denotes the length of the key phrases, which is consistently 2 in this case. In Table 28, the length is extended to 3. The frequency column indicates the extent to which the related phrase is prevalent in the company descriptions.

It could be shown from Table 27 that that the term 'renewable energy' is the most frequently occurring key phrase, which predominant emphasis on renewable energy sources within the sustainability discussions of the companies. The frequency of the terms 'energy company and 'low cost' is seen to be the second most frequent, suggesting a noteworthy focus on company identity within the energy industry and the efficient utilisation of monetary resources. Various terms such as 'natural gas', 'power plant', and 'energy efficiency' are observed with high frequencies, indicating the broad wides of subjects and issues that corporations address in their sustainability statements. The presence of keywords such as "environmental impact," "climate change," and "carbon footprint" within organisations indicates a recognition and understanding of the sustainable consideration of their actions and the wider implications for climate change. Moreover, the use of terms such as 'food waste' and 'supply chain' highlights the complex and diverse aspects of sustainability, which go

beyond energy-related issues to embrace the management of waste and logistical factors.

Table 27 2-gram keywords

keyword	ngram	freq
renewable energy	2	662
natural gas	2	256
solar energy	2	256
technology company	2	240
energy storage	2	237
clean energy	2	236
energy efficiency	2	218
energy company	2	210
low cost	2	209
oil gas	2	204
real time	2	166
solar power	2	163
power plant	2	155
power generation	2	153
high quality	2	145
waste management	2	135
electric vehicle	2	124
supply chain	2	120
wide range	2	117
wind turbine	2	114
energy management	2	109
energy solutions	2	107
energy consumption	2	107
next generation	2	107
fuel cell	2	106
energy system	2	104
solar cell	2	102
water treatment	2	102
energy technology	2	101
research development	2	101

The below table expands the 2-word phrases into 3-word phrases, which are composed of three words and serve as key phrases. It could be shown in
Table 28 that the phrase 'renewable energy firm' holds the highest frequency. It suggests a significant focus on companies that specialise in renewable energy within the sustainability sector. The subsequent terms, namely 'oil natural gas' and 'energy storage system', also have frequencies of 46 and 41 respectively.

Table 28 3-gram keywords

keyword	ngram	freq
renewable energy company	3	53
oil natural gas	3	46
energy storage system	3	41
lithium ion battery	3	40
greenhouse gas emission	3	37
renewable energy project	3	33
oil gas company	3	30
renewable energy source	3	30
oil gas exploration	3	29
renewable energy technology	3	27
oil gas industry	3	27
company renewable energy	3	26
utility scale solar	3	25
renewable energy solutions	3	24
exploration production company	3	23
solar energy system	3	22
gas exploration production	3	21
waste management company	3	20
energy storage technology	3	20
energy technology company	3	20
renewable energy system	3	20
exploration development company	3	19
mineral exploration company	3	19
renewable energy generation	3	18
power purchase agreement	3	17
energy company solar	3	17
clean technology company	3	17
clean renewable energy	3	16
renewable energy industry	3	16
solar cell module	3	16

These numbers indicate a notable emphasis on natural resources and energy storage solutions within the company self-description. The use of terms like 'lithium ion battery' and 'greenhouse gas emission' highlights the significance of advanced energy storage technology and the environmental problems associated with emissions in discussions surrounding sustainability. The repeated occurrence of terminology pertaining to the "oil and gas" sector, such as "oil and gas company," "oil and gas exploration," and "oil and gas industry," highlights the ongoing discourse surrounding natural resource sector within the context of sustainability. The table also presents a wide range of subjects that cover renewable energy projects, sources, technologies, and solutions. This underscores the comprehensive approach to renewable energy

within the context of corporate sustainability discussions. Terms such as 'waste management company and 'clean technology company' underscore the comprehensive nature of sustainability considerations, encompassing waste management and clean technologies.

Compared with the 2-gram keywords, the table including three-word key phrases typically demonstrates a greater degree of precision in the identified concepts. For example, the inclusion of the terms 'energy storage system' and 'lithium ion battery' in the three-word table offers a more detailed understanding of the specific energy technologies, as opposed to the broader term 'energy system' used in the two-word table. Besides, the table of three-word key phrases appears to include a broader range of subjects, incorporating more specific facets of energy technology, exploration, production, and environmental consequences. As a result, it provides a more comprehensive and varied depiction.

5.4.3 Summary of the topic solution

We summarize the outputs from our STM model with the optimal number of topics ($K = 17$) into Table 29 It shows an overview of the 17-topic solution and the corpus-level topic proportions. For each topic, top 7 loading words are shown by calculating the FREX score. we labelled manually labelled each topic in the second column, as well as their corresponding topic proportions (mean of each topic's proportion) across the whole corpus.

Table 29 Labels, distribution and top 7 loading words in the 17-topics solution

#	Topic	Prop (%)	Top 7 loading words
1	Waste Management	5.34	waste, end, tech, collection, management, startup, disposal
2	Water Treatment	5.10	water, air, treatment, wastewater, quality, purification, pollution
3	Sustainable Materials	6.43	material, plastic, economy, sustainable, circular, product, alternative
4	Battery Storage	5.46	battery, storage, cell, manufacturing, generation, high, manufacturer
5	Solar Energy	7.89	solar, electricity, power, grid, developer, photovoltaic, panel
6	Wind Power Generation	2.80	wind, turbine, long, term, structure, small, generator
7	Community Wellbeing	4.53	local, community, city, people, space, food, network
8	Market Operations	6.56	consumer, online, supply, customer, brand, marketplace, chain
9	Client Support	5.05	range, client, wide, expertise, entire, full, team
10	Gas/Oil Exploration	9.14	oil, development, exploration, property, gas, project, mining
11	Energy Efficiency	7.87	energy, renewable, efficiency, consumption, lighting, building, utility
12	Fuel technology	8.40	fuel, heat, research, hydrogen, technology, patent, low
13	Organic Farming	5.77	organic, plant, crop, soil, ingredient, protein, biomass
14	Sustainable Consideration	6.82	climate, impact, sustainability, social, change, carbon, organization
15	Electric Vehicle	3.03	vehicle, electric, car, transportation, infrastructure, charge, electronic
16	Analytics	7.39	software, analytic, data, control, monitoring, cloud, intelligent
17	Retail Solutions	2.41	part, retail, market, solution, new, company, large

As shown in the table, topics extracted from company descriptions could be classified into institutional signals and task signals. The former (14.01%) reveals how organizational characteristics are utilized as signals including Topics 8, 9, 14 and 17 while the latter (85.99%) covers sustainability activities. Based on the top loading words and the original text from company descriptions, Topic 10 (Gas/Oil Exploration) demonstrates how companies utilize more advanced and clean technology than to improve the efficiency than conventional gas/oil extraction and increase the raw material extraction. Topic 12 (Fuel technology) mainly emphasizes the innovations on fuel and hydrogen technology. Topic 5 (Solar Energy) mainly illustrates that solar products or projects are developed by companies as renewable energy to product electricity. The institutional signals demonstrate the quality

indicators that are employed by companies, illustrating the signals from the companies to the investors. Topic 14 (Sustainable Consideration) is the most dominant one among the institutional signals. Sustainable Consideration emphasize companies' strong social conscience and their consideration of social sustainability. It represents companies' awareness of their impact on the environment and how their innovations could reduce the carbon footprint and promote social change.

5.4.4 Marginal effects

As mentioned, document-level covariates could be included in STM to influence the topic solution, which allows us to measure the systematic changes in topic prevalence over the metadata (Roberts et al., 2014). In specific, we construct regression models for each topic where the topic proportion is the dependent variable and document-level covariates from the topic prevalence component (Section 5.3.5) in STM are independent variables. The regression for each topic is as follows:

$$\theta_{k,d} = \beta_1 FundAmount_d + \beta_2 FundRounds_d + \beta_3 CorpYear_d + \beta_4 EmpNum_d + \varepsilon \quad (51)$$

where $k \in (1, \dots, K)$, and θ represents the proportion of topic from topic solution.

We compute the coefficients in small batches by drawing topic proportions from the variational posterior and get the average of the results. Then we are able to capture the changes of topic proportions utilising a marginal effect framework.

Funding amount

Figure 33 represents the marginal effects of the amount of raised funding on the prevalence of signals when each other variable (number of funding rounds, year of corporation and number of employee) is held at its sample median. The dotted line illustrates the zero effect. The distance of each topic from the dotted line shows how much its proportion across the whole corpus changes as the funding amount shift from the smallest (1,000 USD) to the largest (> 15.8 billion USD). The topics at the right-hand side indicate their increases in topic prevalence as the funding amount grow while the topics at the left-hand side demonstrate more popularity within descriptions from companies which raised less funding.

As shown in Figure 33, five topics (on the right-hand side) show increases when we shift from the smallest funding amount to the largest funding amount.

Among them, Energy Efficiency (Topic 11) and Battery Storage (Topic 4) indicate the most significant increase (9% approximately), indicating that these two signals are more dominant in the companies' descriptions with larger amount of funding.

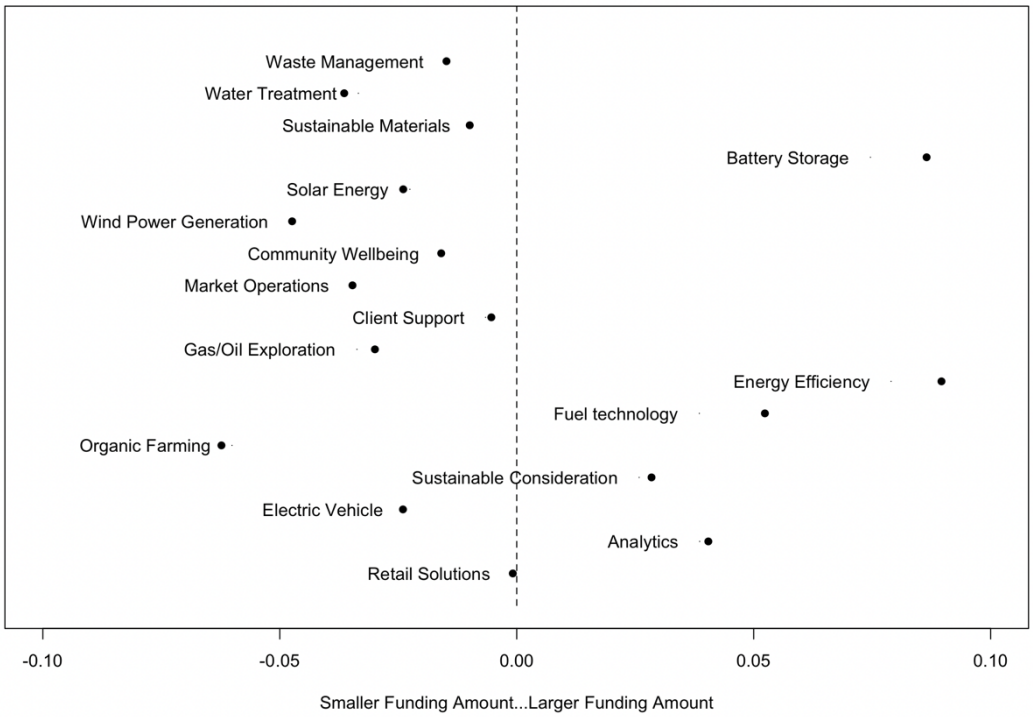

Figure 33 The changes in the expected topic proportions from smaller funding amount to larger funding amount

Three other signals are also adopted popularly by companies which raised larger amount of funding, including Fuel Technology (Topic 12), Analytics (Topic 16) and Sustainable Consideration (Topic 14). Compared with four task signals, the only institutional signal Sustainable Consideration is more likely to be utilised by companies to describe themselves. In contrast, more topics indicate more popularity within descriptions from companies which raised smaller funding amount. For instance, the topic proportion of Organic Farming (Topic 13) from companies which obtain larger amount of funding is approximately 6% lower than that from companies which raised smaller amount of funding, indicating that Organic farming attracts less interest from the large capital investment.

Funding rounds

In the same way, we also estimate the marginal effects of the number of funding rounds on the prevalence of signals when each other variable is held at its sample median. The topics at the right-hand side indicate their increases in topic prevalence as the funding rounds grow while the topics at the left-hand side demonstrate more popularity within descriptions from companies which underwent less funding rounds.

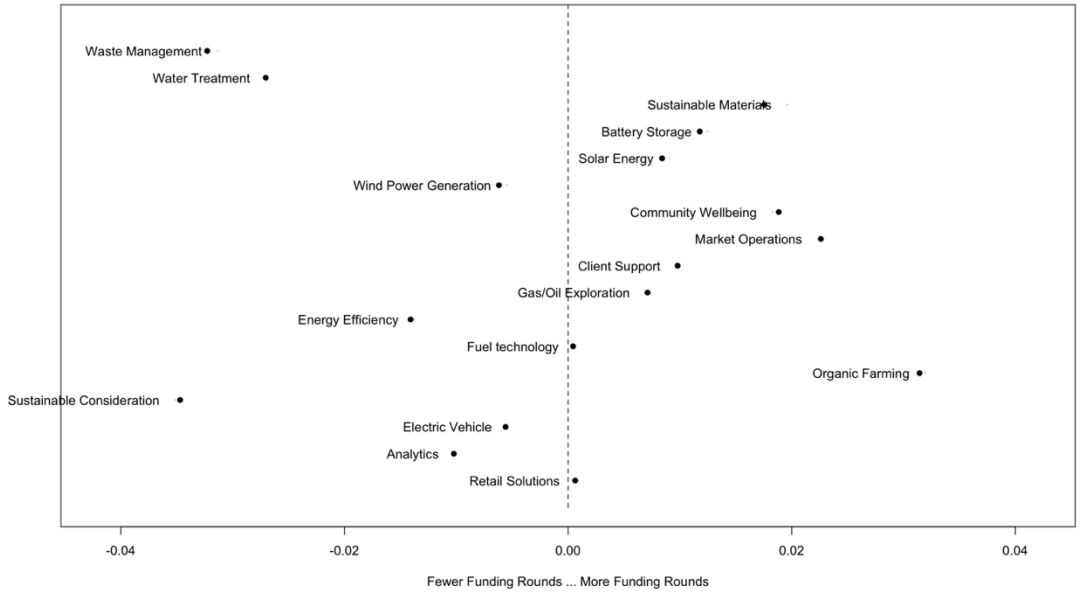

Figure 34 The changes in the expected topic proportions from less funding rounds to more funding rounds

The biggest change occurs to the proportion of Organic Farming topic, increasing over 3% when we shift from the least funding amount to the most funding rounds while the proportion of sustainable consideration topic decreases most. Organic Farming is a popular task signal adopted by companies as it is more dominant within descriptions from companies which went through more funding rounds. In contrast, the Sustainable Consideration topic is more popular from companies that underwent fewer funding rounds. As the number of funding rounds grows, companies tend to de-emphasize their environmental responsibility. There are some distinct changes for task signals including Waste Management and Water Treatment, which owns fewer prevalence as the funding rounds continue increasing while one institutional signal Online Market Operations is more emphasized by companies to shows as their indicators.

Year of corporation

Similarly, we also examine the marginal effects of the number of years of incorporation. On the right side of the dotted line in Figure 35.

Figure 35 The changes in the expected topic proportions from older company to newer company

There are 7 topics that have more proportions in descriptions from newer companies. Among them, there are four task signals including *Waste Management*, *Organic Farming*, *Electric Vehicles* and Analytics which own more popularity within newer companies. Among them, the biggest change in topic proportion occurs for the Waste Management topic. It indicates that the attention of start-ups is increasingly paid to these sustainability activities. The other three are institutional signals, including *Online Market Operations*, *Client Support Services* and *Sustainable Consideration*. It represents that newer companies are more likely to emphasize their quality by these indicators. By contrast, topics at the left-hand side of the dotted line represent the more prevalence among mature companies. The topic describing companies active in *Wind Power Generation* has the biggest decrease in topic proportion representing its decreasing attractiveness for new start-ups as this can also be considered a mature technology with an established innovation stack.

5.4.5 Mapping the dependence between topics

To display the internal relationships of topics within companies' descriptions, we performed a principal component analysis (PCA) among the topic prevalence of 17 topics extracted from the STM result. Figure 36 shows how these signals are correlated in companies' descriptions. The arrows show the correlation among 17 topics. In addition, we selected 1872 companies by limiting the funding rounds greater than 2, which are shown as points distributed around the arrows.

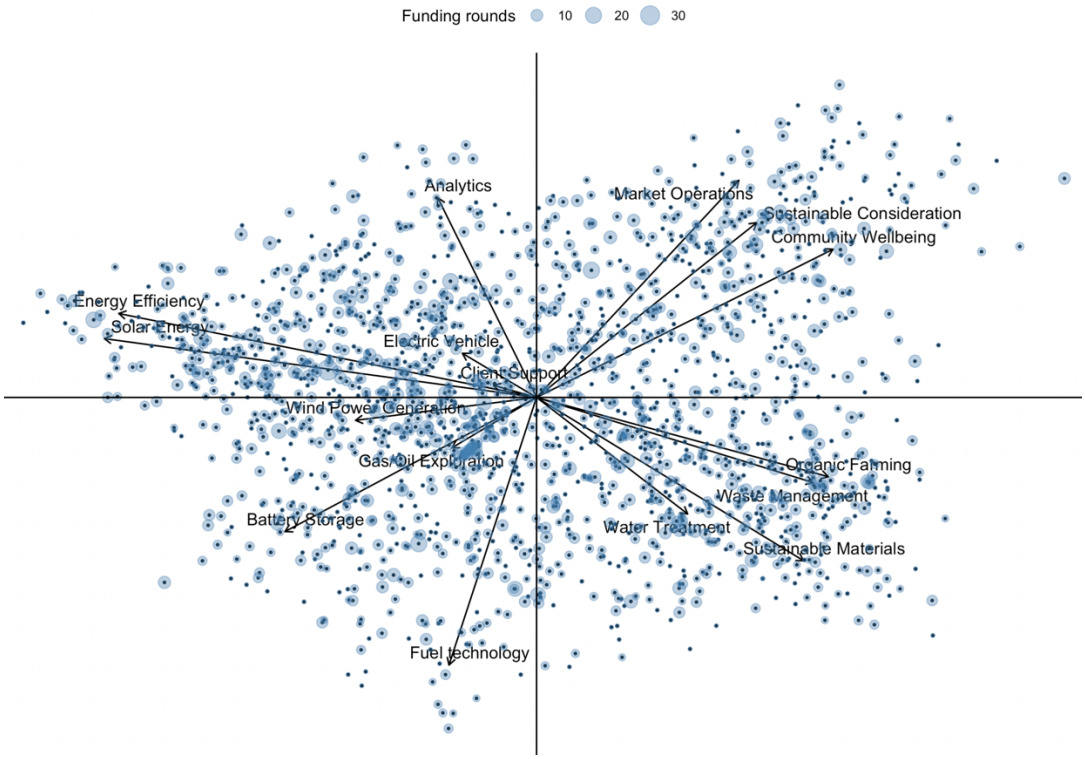

Figure 36 Principal component analysis: topic correlation plots. Point size represents years of incorporation for major companies loading closer to the topic.

As shown in the figure, several clusters of related topics can be observed. For instance, the Solar Energy topic and Energy efficiency are correlated. The energy efficient technologies, such as LED lighting are widely applied in homes or buildings paired with solar panels, decreasing reliance on conventional energy sources, which could contribute to lower energy bills and reduced strain on the power grid. Another three signals including Online Market Operations, Community Wellbeing and Sustainable Consideration also show close connection as Combing community wellbeing and the consideration of sustainability into the online marketing could lead to increased innovation and development of environmentally friendly products and services, which could also promote long-term business success. On the other hand,

there are several pairs of angles of 90 degrees. Analytics and Battery Storage, Market Operations and Water Treatment, Energy efficiency and Fuel Technology as well as Organic Farming and Fuel Technology, which means the is no correlation among each pair of signals. In addition, several cluster of points can also be observed alongside with the signals. For instance, several points are close to two institutional signals (Sustainable Consideration and Market Operations), which represent that companies are more likely to include the two signals to prove their quality to investors.

The graded circle around each point indicates how many funding rounds the company has gone through, ranging from 3 to 34. In general, the percentage of companies which went through over 10 funding rounds is 4.6%, which could be observed Waste Management and Gas/oil Exploration. In contrast, most of them have gone through less than 10 funding rounds. Among them, several company clusters could be observed, representing that more companies are closely distributed with several signals including Gas/oil exploration, Wind power Generation, Electric Vehicle and sustainable consideration.

5.5 Discussion and implications

The present paper contributes to our understanding of the Signalling environment in sustainability-focused ventures and provides recommendations relevant for theory development and investment decision-making. The study demonstrates a long-term pattern of Signalling activities and its investment input implications in firms developing sustainable technologies.

Results suggest that sustainability-focused firms convey information about venture quality, using several latent task and institutional signals. The identified topic modeling solution revealed 17 distinct topics, which were categorized as signals of institutional and task relevance. Task signals related to specific sustainable technologies dominated company descriptions with topic prominence for example in *Solar Energy* and *Energy Efficiency*. On the other hand, institutional signals such as *Online Market* Operations and *Analytics* were also identified, demonstrating significance as quality signals in attracting investment interest. Furthermore, the study's depiction of a correlation between signals suggests that transmission of multiple quality signals enhance Signalling consistency and is effective in attracting investment inputs.

5.5.1 Implications for theory

As an important theoretical framework, Signalling theory has expanded substantially in entrepreneurship (Bafera and Kleinert, 2022). The theory's main premise that ventures portray unobservable qualities with costly, observable activities has been employed to understand various signals and their investment potential (Mochkabadi and Volkmann, 2020).

The present study suggests that latent signals conveyed in self descriptions of ventures associated with sustainable technologies, holds valuable explanatory potential. This is consistent with the importance of rhetorical signals as relevant and timely information that may assist investment decision-making (Steigenberger and Wilhelm, 2018). By applying a topic modeling approach in unstructured data, the study demonstrated how to gain insight into the most salient sustainability signals that companies demonstrate in self-descriptions. More specifically, we extracted two types of signals from self-descriptions: task signals and institutional signals, which indicate how companies utilize signals to indicate their quality and reduce information asymmetry. As such, the study's theoretical contribution are multifaceted, By expanding limited literature to better understand how investment decisions are influenced in sustainability-focused entrepreneurship (Demirel et al., 2019; Wöhler and Haase, 2022) the study echoes the call for theoretical development in entrepreneurship with the introduction of a novel class of computational methods, capitalizing on unstructured data (Berente, Seidel and Safadi, 2019; Hannigan et al., 2019b). In addition, previous studies suggested that some signals could be equally effective during different funding stages (Colombo, 2021) and have unique value over time. To extend this, the study empirically examined the dynamic of signals alongside investment inputs by taking into consideration the total funding amount and funding rounds associated with seed and series funding respectively. Findings suggest that heterogeneity exists between investment outcomes and the prevalence of signals as captured from the topic, indicating the heterogeneous effectiveness of how signals influence the receivers' behaviours.

Finally, several studies have illustrated the interactions among signals and how multiple signals interplay with each other (Drover, Wood and Corbett, 2018; Huang et al., 2022). The study takes another logical step forward to suggest that signal consistency is complimentary to investment inputs in sustainability-focused

entrepreneurship. The PCA approach indicates that certain signals are highly correlated by forming clusters, which suggests the high probability of signals occurring in the same company description. This is an important depiction of internal connections among task and institutional signals.

5.5.2 Implication for practice

The findings of our study bear considerable relevance to practitioners, notably entrepreneurs, startup founders and investors interested in ventures in sustainable technology industries. The study also has significance for policymakers involved in steering a start-up ecosystem in policy priority areas.

First, the study highlights the potential risks emanating from information asymmetries among firms and investors in sustainability. This is a critical assumption, as investors may have limited understanding of the technologies proposed or value propositions developed by start-ups which in turn, may lead to inconsistent investment decision-making. By understanding the fabric of Signalling in the investment process, start-ups can take important steps towards bridging the asymmetry gap and communicate their offerings to potential investors (Handrito, Slabbinck and Vanderstraeten, 2021). Second, our results highlight the increasing importance of institutional Signalling for start-ups seeking early-stage funding. Start-ups with a stronger emphasis on sustainability and social responsibility, may also benefit from highlighting their efforts, which can be achieved through institutional Signalling (Vismara, 2016). Third, our study suggests the important role of task Signalling as a powerful tool for start-ups to differentiate themselves in their market. By professing expertise in areas such as waste management, water treatment, sustainable materials, battery storage, solar energy and energy efficiency, ventures can attract consumer and investor interest.

The present study also has practical significance for investors, suggesting the need to pay close attention to the nature of Signalling efforts manifested by sustainability-focused startups. While there is always some inherent risk associated with investment in sustainable technologies (Cumming, Leboeuf and Schwienbacher, 2017), our study suggests that institutional and task signals may be instrumental to identify ventures with greater potential for success. Investors would usually pay attention to information including product, team and elements of

macroeconomic significance, as factors affecting venture performance Vazirani and Bhattacharjee, 2021; Kaplan et al., 2009). By doing so, they would have an incomplete understanding of the intrafirm dynamics and growth potential which in sustainability focused start-ups tends to have a longer term development dynamic. Instead of emphasizing observable signals conveyed by entrepreneurs (Piva and Rossi-Lamastra, 2018), the present study suggests that latent signals conveyed in self descriptions of ventures associated with sustainable technologies, holds valuable explanatory potential. The use of signals demonstrating online market operations, client support services, sustainable consideration and retail innovation solutions provide quality indicators that may be considered favorably from investors, as these are less vertical to the sustainability domain but centered around it (Schönwälder and Weber, 2023).

Finally, the study has practical value to policy makers looking to nurture sustainability-focused entrepreneurial ecosystems. Our study provides a blueprint for directing investment in areas of interest where there is a seemingly unrelated concomitance between technologies and sustainability focus.

5.6 Limitations and future research recommendations

While the study's innovative depiction of signals through the use of company self-descriptions portray a fruitful avenue for future research in sustainable-focused entrepreneurship, as with all studies, there are several limitations. First, our sample of ventures is drawn from Crunchbase, which may not be representative of all sustainability-focused technology ventures. Future research could address this limitation by constructing samples from multiple sources. In addition, owning to the study's principal focus, only company self-descriptions have been utilized. Future research could expand this line of reasoning by incorporating other sources of signal quality information from unstructured data, such as investor communications, news articles and social media sentiment (Tumasjan, Braun and Stolz, 2021).

Finally, the data analysis is constrained to textual data available and does not take into consideration other influencing factors such as financial metrics. Future research could address this limitation by expanding the scope of investigation to incorporate financial performance metrics to better understand the interplay with task and institutional signals. This is an important depiction that supports this study's assertion

that new ventures can overcome liabilities of newness and smallness and effectively signal to attract investor interest in a rather noisy and opaque Signalling environment.

CHAPTER 6 Conclusions

6.1 Research findings

Study I demonstrates the rating prediction using customer feedback. It summarised a range of methods that has been identified in the literature for improving the predictability of customer feedback, including simple bag-of-words-based approaches and advanced supervised machine learning models, which are designed to work with response variables such as Likert-based rating scores.

In this study, we present a dynamic model that incorporates values from topic membership, an outcome variable from Latent Dirichlet Allocation, with sentiment analysis in an Extreme Gradient Boosting (XGBoost) model used for rating prediction. Sentiment analysis at both document level and sentence level are established in order to detext customer sentiment at both level and to compare how these two levels of sentiment behave distinguish in predicting customer satisfaction. Then the unsupervised approach LDA, as a probabilistic topic model has been applied to extract the latent dimension in the textual data. And then the topic membership which is the outcome variable from LDA could be collaborated with LDA to predict the customer satisfaction.

In order to make the results validated, there are 6 machines learning algorithms (including SVM, NB, KNN, XGBoost, DT, RF) are selected and applied in the classification task. Results indicate that the sentence-level sentiment performs better than the document-level sentiment while underperform the LDA-based topic membership no matter which machine learning classifier is applied. Among these 6 machine learning algorithms, the XGBoost algorithm beat others in terms of the AUC score as well as the relatively high accuracy. Therefore, it is selected to examine the combination for sentiment variable and topic membership as well as the rating prediction task in the next stage.

The results indicate that even incorporating the topic membership from one single topic, the AUC score could be improved. In addition, when the topic membership is applied in the regression for rating prediction, all topics behave the ability to improve the baseline model (using document-level sentiment only) to predict customer ratings. More specially, topic 13 and topic 3 shows the highest contribution to the prediction.

Study II focused on the exploration of online reviews from customers in the OFD domain. This study examines the outcomes of franchising in the service sector from the customer's perspective. This study identifies key issues during the processes of producing and delivering product/services from service providers to customers in service industries. A large dataset is constructed containing 553,807 reviews from 10,894 restaurants (both franchised and independent) posted on *JustEat* and is analysed. At the first stage, this study concentrates on the identification of latent aspects from the unstructured text review content from the word/phrase level using frequency analysis as well as word collocation analysis. By identifying the most occurring nouns and adjective, customers' preference and attitude could be detected intuitionally. What's more, two types of approaches including the statistical approach and linguistic approach are utilised within the collocation analysis in order to identify the relationship between words, thus has a deeper understanding from the phrase level.

At the second stage, a topic modelling approach is employed as the document-level text analysis. The latent dimensions are extracted using STM across the corpus as well as the association between each review and each latent dimension. By linking these latent dimensions to the review metadata, we are able to discover how the review-level covariates affect the topic distribution thus discover the marginal effect of these covariates on how the topics are distributed. Results indicate that for all restaurants, product quality and delivery service are two prominent dimensions mentioned by customers. Besides, topics related to complaint about order processing are more dominant in reviews from franchised restaurants while praise about food quality is more popular from reviews of independent restaurants. We also examine the disparity of performance between franchised and independent businesses by comparing how they perform across each latent dimension between reviews with positive rating and reviews with negative ratings.

The findings could provide managerial insights for franchisors and independent entrepreneurs on what phase of service supply chain could be enhanced, especially in online food delivery.

Study III extends our unstructured data scope to the firm-level data. The need for climate action has increased attention on sustainability-focused entrepreneurship. In this context, entrepreneurial firms play a fundamental role in developing high-technology solutions for decarbonization, but face funding gaps due to liabilities of newness and smallness. Despite the importance of signalling in entrepreneurship,

little is known on what and how to effectively signal to attract investor interest in small ventures that develop sustainable technologies. Study III suggests a topic modelling solution to identify signals presented in company self-descriptions and areas of activity, alongside their investment inputs. Using data extracted from CrunchBase, a corpus of 2,300 self-descriptions of sustainable technologies ventures over a period of 10 years, this study provides novel insights into the signalling environment of sustainability-focused entrepreneurship. Results indicate two types of signals contained in companies' self-descriptions, which can act as latent signals that could signal themselves to the investors. In addition, the association between the signals and the organisation context of the entrepreneurship is also examined through a topic modelling approach by estimating the marginal effect of the organisation context, which could display the heterogeneity of various signals across different organisation context.

Table 30 Research questions and findings

Research questions	**Findings**
RQ1: How can be information from unstructured textual data extracted in business settings?	Despite of the challenge of ncovering information from textual data due to its specific properties, including high dimensionality and sparsity, the thesis could extract useful information present in textual data within business settings by employing hierarchical text mining approaches. This research incorporates appropriate methodologies from related disciplines such as natural language processing, information retrieval, and machine learning
RQ2: What is the role of unstructured data from the business perspective?	The usage of textual data has the capacity to uncover patterns and provide substantial insights to improve the operations management. It can significantly enhance the decision-making process by transforming unstructured data into actionable insights.
RQ3: How can unstructured data be linked with the structured data and the business context?	By incorporating the unstructured data with structured data, the relationship between the qualitative information and quantitative data could be explored either by using quantitative information to influence the qualitive information extraction or employing the value from unstructured data to predict or explain the structured data.

6.2 Theoretical implications

This thesis aims to conduct an extensive analysis for research focused on unstructured data, which contributes significantly to the body of literature on the informational value of unstructured data in business settings. This thesis specifically explores unstructured data from two areas: the service industry and entrepreneurship research. This dissertation discovered how the output from the unstructured data could be linked with the business context, which originates from one of characteristics of text that the meaning of texts is highly related with the specific context. The rich value in the textual data significantly influences consumer behaviour and organizational strategies in these domains.

First, the implication of this dissertation is to build the connection between unstructured data and structured data. The research extends beyond simple conversion and instead establishes a complex connection between unstructured data and structured data from the business. This connection is quantified through an examination of the close relationships between unstructured and structured data. This entails a thorough analysis of the relationship between unstructured data, which contains hidden insights and patterns, and structured data, which is easier to analyse. The purpose is to provide significant insights that are crucial for the success of an organisation. Therefore, the investigation holds significant importance in comprehending how the utilisation of latent information included in unstructured data could be combined or incorporated with the structured data from business (numeric values such as sales). By quantifying the association between both, it demonstrates that the value in the unstructured data that could be collaborated with structured data in the future research to improve business processes or optimise performance.

Second, the text mining techniques presented in this thesis also have significant theoretical implications for the discipline of Natural Language Processing (NLP). This thesis contributes to the more advanced and NLP techniques that can handle the difficulties presented by textual datasets that are high-dimensional, sparse, and noisy. The use of advance text mining techniques contributes to developing more sophisticated NLP models that can effectively reduce dimensionality or handle high-dimensional data without losing critical information. Because of the inherent large vocabulary in natural language, the high dimensionality of textual data is a main problem in traditional NLP techniques. This study has analysed the textual data at

different granularity. By combining different levels of NLP techniques (word level, sentence level and document level), the information contained in textual data is extracted and categorised into hierarchical knowledge which can be easily understood by human beings. Therefore, the high-dimensional textual data could be analysed by the NLP techniques without losing essential information, which is crucial for tasks such as text classification and sentiment analysis. What's more, as mentioned in the literature review, the sparsity problem among the textual data leads to sparse representation of textual data in vector space models. Our hierarchical techniques employed in this thesis contribute to addressing this issue by proposing more robust feature extraction and representation techniques, which are able to capture more semantic similarities. Improving the handling of sparsity in text data directly enhances the performance of NLP techniques. This thesis contributes to the improvement of text mining approaches, which in turn helps in the development of more efficient searching and ranking algorithms capable of handling the complexity of natural language. It is crucial in the era of big data due to the need of accurately retrieving valuable information from enormous textual datasets.

What's more, the thesis employs a data-driven methodology to thoroughly examine the potential of extracting meaningful information from unstructured data, which aims to generate coherent and actionable insights that can greatly enhance business intelligence and inform strategic decision-making. The utilisation of a mixed-methods design in this study is particularly notable due to its ability to permit a comprehensive examination of unstructured data. This thesis proposes a novel model for to analysing unstructured textual data (study I) and proved the excellent ability of the topic modelling approach in understanding the content of text despite the high dimensionality of text by incorporating the latent dimensions from textual content. More specifically, the hierarchical text analysis techniques are applied to understand the text from different levels. The utilisation of exploratory and predictive analysis techniques on vast amounts of unstructured textual data allows for a more thorough investigation of research questions, with an emphasis on several levels of analysis to encompass the diverse characteristics of unstructured data, which enables the investigation of complex patterns, correlations, and structures inherent in the data. From a methodological point of view, it not only increases the comprehensiveness and scope of the examination but also guarantees the strength and dependability of

the results, which can offer an in-depth comprehension of textual data in the corporate environment.

6.3 Practical implications

This research aims to provide in-depth insights and a comprehensive framework that can assist businesses in effectively navigating the challenges posed by these characteristics. By doing so, businesses will be able to fully leverage unstructured textual data to gain a competitive edge in the marketplace. The practical implication of translating research findings into concrete strategies for companies to properly utilise the inherent business value of could be discussed in several aspects.

First, this research could provide insights for firms to maintain deep and effective communication with stakeholders. From the customer side, customer feedback acting as the online interaction with the businesses is a tool to communicate their opinions and thoughts to the merchants. Customer opinions is crucial for organisations to establish service improvement or service recovery. These publicly shared opinions serve as significant sources of information. In order to accomplish this objective, it is highly advised that firms enhance their business analytics skills towards the textual content from customers. The utilisation of substantial textual data could contribute to enhancing service design and delivery, especially in service-oriented organisations by revealing hidden insights and collective sentiments pertaining to products, and services. From the investors and regulatory side, the textual data help companies to demonstrate themselves to investors so that investors can identify market trends for safeguard investments.

Second, gaining a comprehensive understanding of the significance of unstructured data in business setting and its integration with structured data is crucial for enhancing the business intelligence level within organisations. The integration enhances business intelligence frameworks, which are able to allow for more thorough decision-making processes. This integration between two types of data enables a comprehensive understanding of corporate operations, customer behaviours, and market trends by utilising both quantitative structured data and qualitative unstructured insights. As proposed by Abdullah and Ahmad (2015), the unstructured data can be utilised to enhance organisational decision-making. They designed a model to enrich and complement unstructured data with structured data through metadata creation. Therefore, by incorporating diverse data types, the decision-

making process can be enhanced with a more extensive data infrastructure. The multidimensional analysis of unstructured text as well as the integration with quantitative data could provide a more valuable insight base for enterprises (Radhakrishna et al., 2017).

What's more, the thesis could also provide managerial insights for marketer to understand current market trends and develop competitor strategies. The informational value within unstructured data including customer feedback or social medica could enable marketers to stay updated with the latest trends affecting the industry. For example, a sudden surge in discussions about a specific product feature can signal a shift in consumer preferences. The future market trends could be discovered by analysing patterns and also predicted using the textual data, which could enable businesses to be proactive rather than reactive, adjusting their strategies to seize opportunities or mitigate risks. By analyzing the text from competitor press releases, financial reports, and other public communications, companies can gain insights into the strategic moves of competitors, such as new product launches, expansions, or changes in management.

6.4 Limitations and future research

One limitation of this study is that it only focuses on a single type of unstructured data. The research primarily emphasises textual data as the principal type of unstructured data. Even the textual data is most widely discovered in business setting, this dissertation doesn't consider other types such as photos, videos, and audio data. Although textual data offers valuable insights, the omission of other types of unstructured data could potentially restrict the comprehensiveness of the conclusions. The sole dependence on textual data may fail to comprehensively encompass the complicated aspects of consumer experiences, viewpoints, and entrepreneurial signalling, hence perhaps disregarding insights that could be obtained from non-textual data sources.

Another limitation comes from the scope of the thesis. In this thesis, we used the online reviews from OFD to represent the customer-level data and the company self-disclosure to indicate the firm level data. However, there are a broader existence of textual data in the business settings. The generalizability of the conclusions obtained from these particular sources may be limited to other types of text or industries, hence potentially constraining the usefulness of the insights. In order

enhance the generalizability and applicability of the results, subsequent studies must consider the utilisation of diverse data sources across various platforms and industries. The process of examining data from several sources can be advantageous in addressing inherent biases associated with particular platforms and in encompassing a wider range of user behaviours, interests, and experiences. The multiple sources of data have the potential to enhance the comprehension of dynamics across several sectors, thereby allowing the creation of more resilient and adaptable concepts and strategies.

As to the future research, subsequent investigations had to consider the integration of diverse forms of unstructured data, including photographs, videos, and audio, in order to achieve a whole understanding of the subject. The examination of a variety of information forms has the potential to reveal deeper and broad understandings of customer experiences, entrepreneurial signalling, and other relevant domains. The incorporation of several data formats has the potential to enhance the depth of analysis, thus offering an extensive perspective on the phenomena being examined. For example, Pantano et al. (2021) analysed customers' unstructured data (i.e., images) to provide significant competitive advantages by revealing insights into consumer behaviour and emotions. Therefore, informational value from other types of unstructured data could be explored.

What's more, more broad text mining techniques could be employed in discovering the latent information. This thesis examines the ability of two popular topic models while there are other text mining techniques including Topic2Vec, Doc2Vec and BERT, which are based on word embeddings. These techniques could fulfil the mitigate the shortage of topic modelling methods as the latter is a bag-of-word model while in word embeddings individual words are represented as real-valued vectors in a lower-dimensional space and captures inter-word semantics.

References

Aben, T.A.E., van der Valk, W., Roehrich, J.K. and Selviaridis, K. (2021) Managing information asymmetry in public–private relationships undergoing a digital transformation: the role of contractual and relational governance. *International Journal of Operations and Production Management*. 41 (7), 1145–1191. doi:10.1108/IJOPM-09-2020-0675/FULL/PDF.

Abdullah, M.F. and Ahmad, K., 2015, August. Business intelligence model for unstructured data management. In *2015 International Conference on Electrical Engineering and Informatics (ICEEI)* (pp. 473-477). IEEE.

Adnan, K. and Akbar, R. (2019) An analytical study of information extraction from unstructured and multidimensional big data. *Journal of Big Data*. 6 (1), 91. doi:10.1186/s40537-019-0254-8.

Ahlers, G.K.C., Cumming, D., Günther, C. and Schweizer, D. (2015) Signaling in equity crowdfunding. *Entrepreneurship Theory and Practice*. 39 (4), 955–980. doi:10.1111/etap.12157.

Akbari, M. and Hopkins, J.L. (2022) Digital technologies as enablers of supply chain sustainability in an emerging economy. *Operations Management Research*. 15 (3), 689–710. doi:10.1007/S12063-021-00226-8.

Alghamdi, R. and Alfalqi, K. (2015) A survey of topic modeling in text mining. *International Journal of Advanced Computer Science and Applications*. 6 (1). doi:10.14569/IJACSA.2015.060121.

Aliouche, E.H. and Schlentrich, U. (2009) Does Franchising Create Value? An Analysis of the Financial Performance of US Public Restaurant Firms. *International Journal of Hospitality and Tourism Administration*. 10 (2), 93–108. doi:10.1080/15256480902850943.

Al-Natour, S. and Turetken, O. (2020) A comparative assessment of sentiment analysis and star ratings for consumer reviews. *International Journal of Information Management*. 54. doi:10.1016/j.ijinfomgt.2020.102132.

Anandarajan, M., Hill, C. and Nolan, T. (2019) '*Introduction to Text Analytics*', In: *Practical Text Analytics: Maximizing the Value of Text Data*. Springer, pp. 1–11. doi:10.1007/978-3-319-95663-3_1.

Apostolopoulos, N., Chalvatzis, K.J., Liargovas, P.G., Newbery, R. and Rokou, E. (2020) The role of the expert knowledge broker in rural development:

Renewable energy funding decisions in Greece. *Journal of Rural Studies*. 78, 96–106. doi:https://doi.org/10.1016/j.jrurstud.2020.06.015.

Audretsch, D.B., Bönte, W. and Mahagaonkar, P. (2012) Financial signaling by innovative nascent ventures: The relevance of patents and prototypes. *Research Policy*. 41 (8), 1407–1421. doi:10.1016/J.RESPOL.2012.02.003.

Audretsch, D.B. and Keilbach, M. (2007) The Theory of Knowledge Spillover Entrepreneurship. *Journal of Management Studies*. 44 (7), 1242–1254. doi:10.1111/j.1467-6486.2007.00722.x.

Baars, H. and Kemper, H.G. (2008) Management Support with Structured and Unstructured Data—An Integrated Business Intelligence Framework. *Information Systems Management*. 25 (2), 132–148. doi:10.1080/10580530801941058.

Bafera, J. and Kleinert, S. (2022) Signaling Theory in Entrepreneurship Research: A Systematic Review and Research Agenda. *Entrepreneurship Theory and Practice*. 104225872211384. doi:10.1177/10422587221138489.

Bagheri, A., Saraee, M. and De Jong, F. (2013) Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews. *Knowledge-Based Systems*. 52, 201–213.

Balducci, B. and Marinova, D. (2018) Unstructured data in marketing. *Journal of the Academy of Marketing Science*. 46 (4), 557–590. doi:10.1007/s11747-018-0581-x.

Baltacioglu, T., Ada, E., Kaplan, M.D., Yurt And, O. and Cem Kaplan, Y. (2007) A new framework for service supply chains. *The Service Industries Journal*. 27 (2), 105–124.

Banerjee, S. and Chua, A.Y.K. (2016) In search of patterns among travellers' hotel ratings in TripAdvisor. *Tourism Management*. 53, 125–131.

Bank, W. (2021) *United Kingdom: Distribution of gross domestic product (GDP) across economic sectors from 2010 to 2020*. Available at https://www.statista.com/statistics/270372/distribution-of-gdp-across-economic-sectors-in-the-united-kingdom/.

Bapna, S. (2019) Complementarity of signals in early-stage equity investment decisions: Evidence from a randomized field experiment. *Management Science*. 65 (2), 933–952.

Barringer, B.R. and Ireland, R.D. (2010) *Successfully launching new ventures*. New York: Pearson.

Barthélemy, J., Graf, N. and Karaburun, R. (2021) Good but not so great: The impact of chain affiliation on guest satisfaction and guest satisfaction extremeness. *International Journal of Hospitality Management*. 94, 102828. doi:https://doi.org/10.1016/j.ijhm.2020.102828.

Bastani, K., Namavari, H. and Shaffer, J. (2019) Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. *Expert Systems with Applications*. 127, 256–271. doi:10.1016/j.eswa.2019.03.001.

Bates, M. and Weischedel, R.M. (1993) *Challenges in natural language processing.*

Batra, S. and Bawa, S. (2010) Using lsi and its variants in text classification. In: *Advanced Techniques in Computing Sciences and Software Engineering*. Springer. pp. 313–316.

Baum, J.A.C. and Silverman, B.S. (2004) Picking winners or building them? Alliance, intellectual, and human capital as selection criteria in venture financing and performance of biotechnology startups. *Journal of Business Venturing*. 19 (3), 411–436. doi:10.1016/S0883-9026(03)00038-7.

Beckman, C.M., Burton, M.D. and O'Reilly, C. (2007) Early teams: The impact of team demography on VC financing and going public. *Journal of Business Venturing*. 22 (2), 147–173. doi:10.1016/j.jbusvent.2006.02.001.

Bento, N., Gianfrate, G. and Thoni, M.H. (2019) Crowdfunding for sustainability ventures. *Journal of Cleaner Production*. 237, 117751. doi:10.1016/j.jclepro.2019.117751.

Berente, N., Seidel, S. and Safadi, H. (2019) Research Commentary—Data-Driven Computationally Intensive Theory Development. *Information Systems Research*. 30 (1), 50–64. doi:10.1287/isre.2018.0774.

Bergh, D.D., Connelly, B.L., Ketchen, D.J. and Shannon, L.M. (2014) Signalling Theory and Equilibrium in Strategic Management Research: An Assessment and a Research Agenda. *Journal of Management Studies*. 51 (8), 1334–1360. doi:10.1111/joms.12097.

Bergset, L. (2015) The Rationality and Irrationality of Financing Green Start-Ups. *Administrative Sciences*. 5 (4), 260–285. doi:10.3390/admsci5040260.

Bergset, L. and Fichter, K. (2015) Green start-ups – a new typology for sustainable entrepre-neurship and innovation research. *Journal of Innovation Management*. 3 (3), 118–144. doi:10.24840/2183-0606_003.003_0009.

Bholat, D.M., Hansen, S., Santos, P.M. and Schonhardt-Bailey, C. (2015) Text Mining for Central Banks. *SSRN Electronic Journal*. doi:10.2139/SSRN.2624811.

Bilalli, B., Abelló, A., Aluja-Banet, T. and Wrembel, R. (2018) Intelligent assistance for data pre-processing. *Computer Standards and Interfaces*. 57, 101–109. doi:10.1016/J.CSI.2017.05.004.

Birjali, M., Kasri, M. and Beni-Hssane, A. (2021) A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-based systems*. 226, 107134.

Blei, D., Ng, A. and Jordan, M. (2003) Latent dirichlet allocation. *Journal of Machine Learning Research*. 3, 993–1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf?ref=https://githubhelp.com.

Blei, D.M. (2012) Probabilistic topic models. *Communications of the ACM*. 55 (4), 77–84. doi:10.1145/2133806.2133826.

Blei, D.M. and Lafferty, J.D. (2007) A correlated topic model of science. *The annals of applied statistics*. 1 (1), 17–35.

Block, J.H., De Vries, G., Schumann, J.H. and Sandner, P. (2014) Trademarks and venture capital valuation. *Journal of Business Venturing*. 29 (4), 525–542. doi:10.1016/j.jbusvent.2013.07.006.

Bojovic, N., Genet, C. and Sabatier, V. (2018) Learning, signaling, and convincing: The role of experimentation in the business modeling process. *Long Range Planning*. 51 (1), 141–157. doi:10.1016/j.lrp.2017.09.001.

Boon-Itt, S., Wong, C.Y. and Wong, C.W.Y. (2017) Service supply chain management process capabilities: Measurement development. *International Journal of Production Economics*. 193, 1–11.

Breiman, L. (2001) Random forests. *Machine Learning*. 45 (1), 5–32. doi:10.1023/A:1010933404324/METRICS.

Brickley, J.A. and Dark, F.H. (1987) The choice of organizational form the case of franchising. *Journal of Financial Economics*, 18(2), pp.401-420.

Brigl, T., 2018. *Extracting Reliable Topics Using Ensemble Latent Dirichlet Allocation*. Doctoral dissertation. Technische Hochschule Ingolstadt.

Brintrup, A. (2021) AI in the Supply Chain: a classification framework and critical analysis of current state. In: *Oxford Handbook of Supply Chain Management*. OUP USA. doi:10.1093/oxfordhb/9780190066727.013.24.

British Franchised Association (2017) *Turnover of the franchise industry in the United Kingdom (UK) from 1985 to 2015 (in billion GBP)*. Available at https://www.statista.com/statistics/665673/franchise-industry-turnover-united-kingdom-uk/#statisticContainer

British Franchised Association (2018) *2018 Franchise landscape.* Available at https://wzr.ug.edu.pl/fid/upload/files/NatWest_Franchise_Landscape_2018.pdf

Brown, R. and Lee, N. (2019) Strapped for cash? Funding for UK high growth SMEs since the global financial crisis. *Journal of Business Research*. 99, 37–45. doi:10.1016/j.jbusres.2019.02.001.

Brown, R.D. (2013) Selecting and weighting N-grams to identify 1100 languages. In: *Text, Speech, and Dialogue: 16th International Conference*. 2013 pp. 475–483. doi:10.1007/978-3-642-40585-3_60.

Bürer, M.J. and Wüstenhagen, R. (2009) Which renewable energy policy is a venture capitalist's best friend? Empirical evidence from a survey of international cleantech investors. *Energy Policy*. 37 (12), 4997–5006. doi:10.1016/j.enpol.2009.06.071.

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

Cao, J., Xia, T., Li, J., Zhang, Y. and Tang, S. (2009) A density-based method for adaptive LDA model selection. *Neurocomputing*. 72 (7–9), 1775–1781. doi:10.1016/j.neucom.2008.06.011.

Carvell, S.A., Canina, L. and Sturman, M.C. (2016) A comparison of the performance of brand-affiliated and unaffiliated hotel properties. *Cornell Hospitality Quarterly*. 57 (2), 193–201.

Cavallo, A., Ghezzi, A., Colombelli, A. and Casali, G.L. (2020) Agglomeration dynamics of innovative start-ups in Italy beyond the industrial district era. *International Entrepreneurship and Management Journal*. 16 (1), 239–262. doi:10.1007/s11365-018-0521-8.

Chan, C.S.R., Parhankangas, A., Sahaym, A. and Oo, P. (2020) Bellwether and the herd? Unpacking the u-shaped relationship between prior funding and

subsequent contributions in reward-based crowdfunding. *Journal of Business Venturing*. 35 (2), 105934. doi:10.1016/J.JBUSVENT.2019.04.002.

Chang, K.C. and Cheng, Y.S. (2021) How online service recovery reviews influence behavioral intentions in the hospitality context: Regulatory focus and loss aversion perspectives. *Journal of Hospitality and Tourism Management*. 46, 440–455. doi:10.1016/J.JHTM.2021.01.014.

Chatterjee, S., Goyal, D., Prakash, A. and Sharma, J. (2021) Exploring healthcare/health-product ecommerce satisfaction: A text mining and machine learning application. *Journal of Business Research*. 131, 815–825.

Chen, J., Heng, C.S., Tan, B.C.Y. and Lin, Z. (2018) The distinct signaling effects of R&D subsidy and non-R&D subsidy on IPO performance of IT entrepreneurial firms in China. *Research Policy*. 47 (1), 108–120. doi:10.1016/j.respol.2017.10.004.

Chen, T. and Guestrin, C. (2016) Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016 pp. 785–794. doi:10.1145/2939672.2939785.

Chen, Y., Wang, W., Liu, Z. and Lin, X. (2009) Keyword search on structured and semi-structured data. *SIGMOD-PODS'09 - Proceedings of the International Conference on Management of Data and 28th Symposium on Principles of Database Systems*. 1005–1009. doi:10.1145/1559845.1559966.

Chen, Y.-M. and Su, C.-T. (2014) Brand equity heterogeneity among strategic groups in service franchising. *The Service Industries Journal*. 34 (9–10), 867–884.

Cheng, Z., Ding, Y., Zhu, L. and Kankanhalli, M. (2018) Aspect-aware latent factor model: Rating prediction with ratings and reviews. In: *Proceedings of the 2018 world wide web conference*. pp. 639–648.

Chevalier, J.A. and Mayzlin, D. (2006) The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*. 43 (3), 345–354.

Chong, A.Y.L., Ch'ng, E., Liu, M.J. and Li, B. (2017) Predicting consumer product demands via Big Data: the roles of online promotional marketing and online reviews. *International Journal of Production Research*. 55 (17), 5142–5156. doi:10.1080/00207543.2015.1066519.

Chung, S., Singh, H. and Lee, K. (2000) Complementarity, Status Similarity and Social Capital as Drivers of Alliance Formation. *Strategic Management Journal*. 21, 1–22.

Clemons, E.K., Gao, G.G. and Hitt, L.M. (2006) When Online Reviews Meet Hyperdifferentiation: A Study of the Craft Beer Industry. *Journal of Management Information Systems*. 23 (2), 149–171. doi:10.2753/MIS0742-1222230207.

Cohen, L., Manion, L. and Morrison, K. (2017) *Research methods in education*. London: Routledge.

Colombo, M.G., Meoli, M. and Vismara, S. (2019) Signaling in science-based IPOs: The combined effect of affiliation with prestigious universities, underwriters, and venture capitalists. *Journal of Business Venturing*. 34 (1), 141–177. doi:10.1016/J.JBUSVENT.2018.04.009.

Colombo, O. (2021) The Use of Signals in New-Venture Financing: A Review and Research Agenda. *Journal of Management*. 47 (1), 237–259. doi:10.1177/0149206320911090.

Combs, J.G., Ketchen Jr, D.J., Shook, C.L. and Short, J.C. (2011) Antecedents and consequences of franchising: Past accomplishments and future challenges. *Journal of Management*. 37 (1), 99–126.

Connelly, B.L., Certo, S.T., Ireland, R.D. and Reutzel, C.R. (2011) Signaling Theory: A Review and Assessment. *Journal of Management*. 37 (1), 39–67. doi:10.1177/0149206310388419.

Cooke, P. (2008) Cleantech and an analysis of the platform nature of life sciences: Further reflections upon platform policies. *European Planning Studies*. 16 (3), 375–393. doi:10.1080/09654310801939672.

Correa, J.C., Garzón, W., Brooker, P., Sakarkar, G., Carranza, S.A., Yunado, L. and Rincón, A. (2019) Evaluation of collaborative consumption of food delivery services through web mining techniques. *Journal of Retailing and Consumer Services*. 46, 45–50.

Costa, V.G. and Pedreira, C.E. (2022) Recent advances in decision trees: an updated survey. *Artificial Intelligence Review 2022 56:5*. 56 (5), 4765–4800. doi:10.1007/S10462-022-10275-5.

Courtney, C., Dutta, S. and Li, Y. (2017) Resolving Information Asymmetry: Signaling, Endorsement, and Crowdfunding Success. *Entrepreneurship Theory and Practice*. 41 (2), 265–290. doi:10.1111/etap.12267.

Cowan, K. and Guzman, F. (2020) How CSR reputation, sustainability signals, and country-of-origin sustainability reputation contribute to corporate brand

performance: An exploratory study. *Journal of Business Research*. 117, 683–693. doi:10.1016/J.JBUSRES.2018.11.017.

Criscuolo, P., Nicolaou, N. and Salter, A. (2012) The elixir (or burden) of youth? Exploring differences in innovation between start-ups and established firms. *Research Policy*. 41 (2), 319–333. doi:10.1016/j.respol.2011.12.001.

Cumming, D.J., Leboeuf, G. and Schwienbacher, A. (2017) Crowdfunding cleantech. *Energy Economics*. 65, 292–303. doi:10.1016/J.ENECO.2017.04.030.

Cunha, F.A.F. de S., Meira, E. and Orsato, R.J. (2021) Sustainable finance and investment: Review and research agenda. *Business Strategy and the Environment*. 30 (8), 3821–3838. doi:10.1002/BSE.2842.

Daft, R.L. and Lengel, R.H. (1986) Organizational Information Requirements, Media Richness and Structural Design. *Management Science*. 32 (5), 554–571. doi:10.1287/mnsc.32.5.554.

Dahlmann, F. and Roehrich, J.K. (2019) Sustainable supply chain management and partner engagement to manage climate change information. *Business Strategy and the Environment*. 28 (8), 1632–1647. doi:10.1002/bse.2392.

Dai, W., Xue, G.R., Yang, Q. and Yu, Y., 2007, July. Transferring naive bayes classifiers for text classification. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence.* pp. 540-545.

Dalle, J., Besten, M. Den and Menon, C. (2017) *Using Crunchbase for economic and managerial research*. OECD Publishing. doi:https://doi.org/10.1787/6c418d60-en.

Danowski, J.A. and Park, D.W. (2009) Networks of the dead or alive in cyberspace: public intellectuals in the mass and internet media. *New Media and Society*. 11 (3), 337–356. doi:10.1177/1461444808101615.

Dant, R.P. (2008) A futuristic research agenda for the field of franchising. *Journal of Small Business Management*. 46 (1), 91–98.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R. (1990) Indexing by latent semantic analysis. *Journal of the American Society for Information Science*. 41 (6), 391–407. doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

Demirel, P., Li, Q.C., Rentocchini, F. and Tamvada, J.P. (2019) Born to be green: new insights into the economics and management of green entrepreneurship. *Small Business Economics*. 52 (4), 759–771. doi:10.1007/S11187-017-9933-Z.

Deveaud, R., SanJuan, E. and Bellot, P. (2014) Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique*. 17 (1), 61–84. doi:10.3166/DN.17.1.61-84.

Dey, A., Jenamani, M. and Thakkar, J.J. (2018) Senti-N-Gram : An n -gram lexicon for sentiment analysis. *Expert Systems with Applications*. 103, 92–105. doi:10.1016/j.eswa.2018.03.004.

Dixon, R.M. and Aikhenvald, A.Y. eds. (2003) *Word: A cross-linguistic typology*. Cambridge University Press.

Donthu, N., Kumar, S., Pandey, N., Pandey, N. and Mishra, A. (2021) Mapping the electronic word-of-mouth (eWOM) research: A systematic review and bibliometric analysis. *Journal of Business Research*. 135, 758–773. doi:10.1016/J.JBUSRES.2021.07.015.

Drover, W., Wood, M.S. and Corbett, A.C. (2018) Toward a Cognitive View of Signalling Theory: Individual Attention and Signal Set Interpretation. *Journal of Management Studies*. 55 (2), 209–231. doi:10.1111/joms.12282.

Duan, L. and Xiong, Y. (2015) Big data analytics and business analytics. *Journal of Management Analytics*. 2 (1), 1–21. doi:10.1080/23270012.2015.1020891.

Dutta, D. and Bose, I. (2015) Managing a Big Data project: The case of Ramco Cements Limited. *International Journal of Production Economics*. 165, 293–306. doi:10.1016/J.IJPE.2014.12.032.

Eberendu, A.C. (2016) Unstructured Data: an overview of the data of Big Data. *International Journal of Computer Trends and Technology*. 38 (1), 46–50. doi:10.14445/22312803/IJCTT-V38P109.

Ebrahimi, S. and Toosi, F.L. (2013) An Analysis of English Translation of Collocations in Sa'di's Orchard: A Comparative Study. *Theory and Practice in Language Studies*. 3 (1). doi:10.4304/tpls.3.1.82-87.

Eddleston, K.A., Ladge, J.J., Mitteness, C. and Balachandra, L. (2016) Do you See what I See? Signaling Effects of Gender and Firm Characteristics on Financing Entrepreneurial Ventures. *Entrepreneurship Theory and Practice*. 40 (3), 489–514. doi:10.1111/etap.12117.

Eggers, F. (2020) Masters of disasters? Challenges and opportunities for SMEs in times of crisis. *Journal of Business Research*. 116, 199–208. doi:10.1016/J.JBUSRES.2020.05.025.

Eggly, S., Penner, L.A., Greene, M., Harper, F.W.K., Ruckdeschel, J.C. and Albrecht, T.L. (2006) Information seeking during “bad news” oncology interactions: Question asking by patients and their companions. *Social Science and Medicine*. 63 (11), 2974–2985. doi:https://doi.org/10.1016/j.socscimed.2006.07.012.

Elitzur, R. and Gavious, A. (2003) Contracting, signaling, and moral hazard: a model of entrepreneurs, ‘angels,’ and venture capitalists. *Journal of Business Venturing*. 18 (6), 709–725. doi:10.1016/S0883-9026(03)00027-2.

Ellram, L.M., Tate, W.L. and Billington, C. (2004) Understanding and managing the services supply chain. *Journal of Supply Chain Management*. 40 (3), 17–32.

Elshakankery, K. and Ahmed, M.F. (2019) HILATSA: A hybrid Incremental learning approach for Arabic tweets sentiment analysis. *Egyptian Informatics Journal*. 20 (3), 163–171.

Enz, C.A., Peiró-Signes, Á. and Segarra-Oña, M.-V. (2014) How fast do new hotels ramp up performance? *Cornell Hospitality Quarterly*. 55 (2), 141–151.

Fang, B., Ye, Q., Kucukusta, D. and Law, R. (2016) Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics. *Tourism Management*. 52, 498–506. doi:10.1016/J.TOURMAN.2015.07.018.

Farkhod, A., Abdusalomov, A., Makhmudov, F. and Cho, Y.I. (2021) LDA-Based Topic Modeling Sentiment Analysis Using Topic/Document/Sentence (TDS) Model. *Applied Sciences*. 11 (23), 11091.

Farrell, M., Wallis, N.C. and Evans, M.T. (2007) A Replication Study of Priorities and Attitudes of Two Nursing Programs’ Communities of Interest: An Appreciative Inquiry. *Journal of Professional Nursing*. 23 (5), 267–277. doi:https://doi.org/10.1016/j.profnurs.2007.01.023.

Felgueiras, M., Batista, F. and Carvalho, J.P. (2020) Creating Classification Models from Textual Descriptions of Companies Using Crunchbase. In: *Communications in Computer and Information Science*. Springer. pp. 695–707. doi:10.1007/978-3-030-50146-4_51.

Ferrati, F. and Muffatto, M. (2020) Using Crunchbase for research in Entrepreneurship: data content and structure. In: *19th European Conference on Research Methodology for Business and Management Studies*. pp. 342–351.

Fico, F.G., Lacy, S. and Riffe, D. (2008) A Content Analysis Guide for Media Economics Scholars. *Journal of Media Economics*. 21 (2), 114–130. doi:10.1080/08997760802069994.

Fisher, G. and Neubert, E. (2022) Evaluating Ventures Fast and Slow: Sensemaking, Intuition, and Deliberation in Entrepreneurial Resource Provision Decisions. *Entrepreneurship Theory and Practice*. 2022 (0), 104225872210932. doi:10.1177/10422587221093291.

Freeman, M. (2019) Can technology innovation save us from climate change? *Journal of International Affairs*. 73 (1), 171–182. doi:https://www.jstor.org/stable/26872787.

Fuller, C.M., Biros, D.P. and Delen, D. (2011) An investigation of data and text mining methods for real world deception detection. *Expert Systems with Applications*. 38 (7), 8392–8398. doi:https://doi.org/10.1016/j.eswa.2011.01.032.

Futagami, K., Fukazawa, Y., Kapoor, N. and Kito, T., 2021. Pairwise acquisition prediction with SHAP value interpretation. *The Journal of Finance and Data Science*, *7*, pp.22-44.

Gaddy, B.E., Sivaram, V., Jones, T.B. and Wayman, L. (2017) Venture Capital and Cleantech: The wrong model for energy innovation. *Energy Policy*. 102, 385–395. doi:10.1016/j.enpol.2016.12.035.

Gans, J., Hsu, D. and Stern, S. (2000) *When Does Start-Up Innovation Spur the Gale of Creative Destruction?* doi:10.3386/w7851.

Geetha, M., Singha, P. and Sinha, S. (2017) Relationship between customer sentiment and online customer ratings for hotels-An empirical analysis. *Tourism management*. 61, 43–54.

Gelbukh, A., Sidorov, G., Han, S.Y. and Hernández-Rubio, E. (2004) Automatic enrichment of a very large dictionary of word combinations on the basis of dependency formalism. *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*. 2972, 430–437. doi:10.1007/978-3-540-24694-7_44.

George, G., Merrill, R.K. and Schillebeeckx, S.J.D. (2021) Digital Sustainability and Entrepreneurship: How Digital Innovations Are Helping Tackle Climate Change and Sustainable Development. *Entrepreneurship: Theory and Practice*. 45 (5), 999–1027. doi:10.1177/1042258719899425.

Ghasemaghaei, M., Eslami, S.P., Deal, K. and Hassanein, K. (2018) Reviews' length and sentiment as correlates of online reviews' ratings. *Internet Research*. 28 (3), 544–563. doi:10.1108/IntR-12-2016-0394.

Ghiassi, M. and Lee, S. (2018) A domain transferable lexicon set for Twitter sentiment analysis using a supervised machine learning approach. *Expert Systems with Applications*. 106, 197–216.

Ghose, A. and Ipeirotis, P.G. (2011) Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE transactions on knowledge and data engineering*. 23 (10), 1498–1512.

Ghosh, S., Lee, L.H. and Ng, S.H. (2015) Bunkering decisions for a shipping liner in an uncertain environment with service contract. *European Journal of Operational Research*. 244 (3), 792–802.

Giannakas, F., Troussas, C., Krouska, A., Sgouropoulou, C. and Voyiatzis, I. (2021) XGBoost and deep neural network comparison: The case of teams' performance. In: *International Conference on Intelligent Tutoring Systems*. 2021 Springer. pp. 343–349.

Giannakis, M. (2011) Conceptualizing and managing service supply chains. *The Service Industries Journal*. 31 (11), 1809–1823.

Gillis, W.E., Combs, J.G. and Yin, X. (2020) Franchise management capabilities and franchisor performance under alternative franchise ownership strategies. *Journal of Business Venturing*. 35 (1).

Gimenez-Fernandez, E.M., Sandulli, F.D. and Bogers, M. (2020) Unpacking liabilities of newness and smallness in innovative start-ups: Investigating the differences in innovation performance between new and older small firms. *Research Policy*. 49 (10), 104049. doi:10.1016/J.RESPOL.2020.104049.

Giones, F. and Francesc, M. (2015) Do actions matter more than resources? A signalling theory perspective on the technology entrepreneurship process. *Technology Innovation Management Review*. 5 (3), 39–45. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2602295.

Gladysz, B. and Kluczek, A. (2017) A framework for strategic assessment of far-reaching technologies: A case study of Combined Heat and Power technology. *Journal of Cleaner Production*. 167, 242–252. doi:10.1016/J.JCLEPRO.2017.08.175.

Griffiths, T.L. and Steyvers, M. (2002) A probabilistic approach to semantic representation. In: *Proceedings of the Twenty-Fourth Annual Conference of Cognitive Science Society*. 2002 pp. 381–386.

Griffiths, T.L. and Steyvers, M. (2004) Finding scientific topics. In: *Proceedings of the National academy of Sciences*. 2004 pp. 5228–5235.

Griffiths, T.L., Steyvers, M. and Tenenbaum, J.B. (2007) Topics in semantic representation. *Psychological Review*. 114 (2), 211–244. doi:10.1037/0033-295X.114.2.211.

Guo, Y., Barnes, S.J. and Jia, Q. (2017) Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management*. 59, 467–483. doi:10.1016/j.tourman.2016.09.009.

Ha, J. and Jang, S.S. (2010) Effects of service quality and food quality: The moderating role of atmospherics in an ethnic restaurant segment. *International Journal of Hospitality Management*. 29 (3), 520–529.

Hallencreutz, J. and Parmler, J., 2021. Important drivers for customer satisfaction–from product focus to image and service quality. *Total Quality Management and Business Excellence*, *32*(5-6), pp.501-510.

Handrito, R.P., Slabbinck, H. and Vanderstraeten, J. (2021) Being pro-environmentally oriented SMEs: Understanding the entrepreneur's explicit and implicit power motives. *Business Strategy and the Environment*. 30 (5), 2241–2254. doi:10.1002/bse.2741.

Hannigan, T.R., Haans, R.F.J., Vakili, K., Tchalian, H., Glaser, V.L., Wang, M.S., Kaplan, S. and Jennings, P.D. (2019a) Topic Modeling in Management Research: Rendering New Theory from Textual Data. *Academy of Management Annals*. 13 (2), 586–632. doi:10.5465/annals.2017.0099.

Hannigan, T.R., Haans, R.F.J., Vakili, K., Tchalian, H., Glaser, V.L., Wang, M.S., Kaplan, S. and Jennings, P.D. (2019b) Topic Modeling in Management Research: Rendering New Theory from Textual Data. *Academy of Management Annals*. 13 (2), 586–632. doi:10.5465/annals.2017.0099.

Harasheh, M. (2022a) Does it Make You Better Off? Initial Public Offerings (IPOs) and Corporate Sustainability Performance: Empirical Evidence. *Global Business Review*. 23 (6), 1375–1387. doi:10.1177/09721509221126851.

Harasheh, M. (2022b) Freshen up before going public: Do environmental, social, and governance factors affect firms' appearance during the initial public offering? *Business Strategy and the Environment*. doi:10.1002/bse.3261.

Hargittai, E., 2006. Hurdles to information seeking: Spelling and typographical mistakes during users' online behavior. *Journal of the Association for Information Systems*, *7*(1), 52-67. doi:10.17705/1jais.00076.

Harrer, T. and Owen, R. (2022) Reducing early-stage Cleantech funding gaps: an exploration of the role of Environmental Performance Indicators. *International Journal of Entrepreneurial Behavior and Research*. 28 (9), 268–288. doi:10.1108/IJEBR-10-2021-0849.

Hart, S.L. and Milstein, M.B. (1999) Global sustainability and the creative destruction of industries. *Sloan Management Review*. 41 (1), 23–33.

Hegeman, P.D. and Sørheim, R. (2021) Why do they do it? Corporate venture capital investments in cleantech startups. *Journal of Cleaner Production*. 294. doi:10.1016/J.JCLEPRO.2021.126315.

Himelboim, I., McCreery, S. and Smith, M. (2013) Birds of a Feather Tweet Together: Integrating Network and Content Analyses to Examine Cross-Ideology Exposure on Twitter. *Journal of Computer-Mediated Communication*. 18 (2), 154–174. doi:10.1111/jcc4.12001.

Hirsimaki, T., Pylkkonen, J. and Kurimo, M. (2009) Importance of high-order N-gram models in morph-based speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*. 17 (4), 724–732. doi:10.1109/TASL.2008.2012323.

Hoenig, D. and Henkel, J. (2015) Quality signals? The role of patents, alliances, and team experience in venture capital financing. *Research Policy*. 44 (5), 1049–1064.

Hofmann, T. (2001) Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*. 42 (1), 177–196.

Horne, J. and Fichter, K. (2022) Growing for sustainability: Enablers for the growth of impact startups – A conceptual framework, taxonomy, and systematic literature review. *Journal of Cleaner Production*. 349. doi:10.1016/J.JCLEPRO.2022.131163.

Houvardas, J. and Stamatatos, E. (2006) N-Gram Feature Selection for Authorship Identification. In: *N-gram feature selection for authorship identification.*

In International conference on artificial intelligence: Methodology, systems, and applications. Springer Verlag. pp. 77–86. doi:10.1007/11861461_10.

Howell, K.E. (2012) *An Introduction to the Philosophy of Methodology*. SAGE Publications Ltd.

Hsu, D.H. (2007) Experienced entrepreneurial founders, organizational capital, and venture capital funding. *Research Policy*. 36 (5), 722–741. doi:10.1016/J.RESPOL.2007.02.022.

Hsu, D.H. and Ziedonis, R.H. (2013) Resources as dual sources of advantage: Implications for valuing entrepreneurial-firm patents. *Strategic Management Journal*. 34 (7), 761–781. doi:10.1002/SMJ.2037.

Hsu, L.-T. (Jane) and Jang, S. (Shawn) (2009) Effects of restaurant franchising: Does an optimal franchise proportion exist? *International Journal of Hospitality Management*. 28 (2), 204–211. doi:https://doi.org/10.1016/j.ijhm.2008.07.002.

Hu, M. and Liu, B. (2004) Mining opinion features in customer reviews. In: *Proceedings of the 19th national conference on Artifical intelligence*. 2004 pp. 755–760.

Hu, N., Koh, N.S. and Reddy, S.K. (2014) Ratings lead you to the product, reviews help you clinch it? The mediating role of online review sentiments on product sales. *Decision Support Systems*. 57, 42–53.

Hu, N., Zhang, T., Gao, B. and Bose, I. (2019) What do hotel customers complain about? Text analysis using structural topic model. *Tourism Management*. 72, 417–426. doi:10.1016/j.tourman.2019.01.002.

Huang, S., Pickernell, D., Battisti, M. and Nguyen, T. (2022) Signalling entrepreneurs' credibility and project quality for crowdfunding success: cases from the Kickstarter and Indiegogo environments. *Small Business Economics*. 58 (4), 1801–1821. doi:10.1007/s11187-021-00477-6.

Hunneman, A., Verhoef, P.C. and Sloot, L.M. (2015) The impact of consumer confidence on store satisfaction and share of wallet formation. *Journal of Retailing*. 91 (3), 516–532.

Hutchineson, S.A. (2004) Education and Grounded Theory. In: *Qualitative Research In Education*. P122-139. Routledge.

Ignatow, G. and Rada Mihalcea (2017) *An introduction to text mining: Research design, data collection, and analysis*. Sage Publications.

Inmon, W.H. and Linstedt, D. (2015) Corporate Data. In: *Data Architecture: a Primer for the Data Scientist*. Morgan Kaufmann. pp. 1–7. doi:10.1016/B978-0-12-802044-9.00001-5.

Inmon, W.H. and Nesavich, A. (2007) *Tapping into unstructured data: Integrating unstructured data and textual analytics into business intelligence*. Pearson Education.

Islam, M., Fremeth, A. and Marcus, A. (2018) Signaling by early stage startups: US government research grants and venture capital funding. *Journal of Business Venturing*. 33 (1), 35–51. doi:10.1016/J.JBUSVENT.2017.10.001.

Isson, J.P. and Harriott, J. (2012) *Win with advanced business analytics: Creating business value from your data*. John Wiley and Sons.

Isson Paul Jean (2018) *Unstructured data analytics: how to improve customer acquisition, customer retention, and fraud detection and prevention.* John Wiley and Sons.

Jain, N., Hasija, S. and Popescu, D.G. (2013) Optimal Contracts for Outsourcing of Repair and Restoration Services. *Operations Research*. 61 (6), 1295–1311. doi:10.1287/opre.2013.1210.

Jolink, A. and Niesten, E. (2021) Credibly reducing information asymmetry: Signaling on economic or environmental value by environmental alliances. *Long Range Planning*. 54 (4), 101996. doi:10.1016/j.lrp.2020.101996.

Jr, R.G. and Unrau, Y. (2010) *Social work research and evaluation: Foundations of evidence-based practice*. Oxford University Press.

Kanaris, I., Kanaris, K., Houvardas, I. and Stamatatos, E., 2007. Words versus character n-grams for anti-spam filtering. *International Journal on Artificial Intelligence Tools*, *16*(06), pp.1047-1067.

Kang, E. and Lam, N.B. (2023) The impact of environmental disclosure on initial public offering underpricing: Sustainable development in Singapore. *Corporate Social Responsibility and Environmental Management*. 30 (1), 119–133. doi:10.1002/csr.2342.

Kapoor, R. and Klueter, T. (2015) Decoding the adaptability-rigidity puzzle: Evidence from pharmaceutical incumbents' pursuit of gene therapy and monoclonal antibodies. *Academy of Management Journal*. 58 (4), 1180–1207. doi:10.5465/AMJ.2013.0430.

Karl Pearson (1904) *On the theory of contingency and its relation to association and normal correlation*. London, Dulau and Co.

Karmiani, D., Kazi, R., Nambisan, A., Shah, A. and Kamble, V. (2019) Comparison of Predictive Algorithms: Backpropagation, SVM, LSTM and Kalman Filter for Stock Market. *Proceedings - 2019 Amity International Conference on Artificial Intelligence, AICAI 2019*. 228–234. doi:10.1109/AICAI.2019.8701258.

Kaufmann, P.J. (1999) Franchising and the choice of self-employment. *Journal of Business Venturing*. 14 (4), 345–362.

Keller, K.L. and Lehmann, D.R. (2003) How do brands create value? *Marketing Management*. 12 (3), 26.

Khanam, Z., Alwasel, B.N., Sirafi, H. and Rashid, M. (2021) Fake news detection using machine learning approaches. In: *IOP Conference Series: Materials Science and Engineering*. 2021 IOP Publishing. p. 12040.

Kidwell, R.E., Nygaard, A. and Silkoset, R. (2007) Antecedents and effects of free riding in the franchisor–franchisee relationship. *Journal of Business Venturing*. 22 (4), 522–544.

Kim, S., Park, H. and Lee, J. (2020) Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis. *Expert Systems with Applications*. 152, 113401. doi:10.1016/j.eswa.2020.113401.

Kim, W.G., Lim, H. and Brymer, R.A. (2015) The effectiveness of managing social media on hotel performance. *International Journal of Hospitality Management*. 44, 165–171. doi:10.1016/J.IJHM.2014.10.014.

Kimjeon, J. and Davidsson, P. (2022) External Enablers of Entrepreneurship: A Review and Agenda for Accumulation of Strategically Actionable Knowledge. *Entrepreneurship Theory and Practice*. 46 (3), 643–687.

Ko, D. G., Mai, F., Shan, Z. and Zhang, D. (2019). Operational efficiency and patient-centered health care: A view from online physician reviews. *Journal of Operations Management*, 65, 353-379.

Ko, E.-J. and McKelvie, A. (2018) Signaling for more money: The roles of founders' human capital and investor prominence in resource acquisition across different stages of firm development. *Journal of Business Venturing*. 33 (4), 438–454.

Koh, Y., Lee, S. and Boo, S. (2009) Does franchising help restaurant firm value? *International Journal of Hospitality Management*. 28 (2), 289–296.

Koltcov, S., Koltsova, O. and Nikolenko, S. (2014) Latent dirichlet allocation: stability and applications to studies of user-generated content. In: *Proceedings of the 2014 ACM conference on Web science*. 2014 pp. 161–165. doi:10.1145/2615569.2615680.

Konuk, F.A. (2019) The influence of perceived food quality, price fairness, perceived value and satisfaction on customers' revisit and word-of-mouth intentions towards organic food restaurants. *Journal of Retailing and Consumer Services*. 50, 103–110.

Korfiatis, N., Stamolampros, P., Kourouthanassis, P. and Sagiadinos, V. (2019a) Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews. *Expert Systems with Applications*. 116, 472–486. doi:10.1016/j.eswa.2018.09.037.

Korfiatis, N., Stamolampros, P., Kourouthanassis, P. and Sagiadinos, V. (2019b) Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews. *Expert Systems with Applications*. 116, 472–486.

Kotsiantis, S.B. (2013) Decision trees: A recent overview. *Artificial Intelligence Review*. 39 (4), 261–283. doi:10.1007/S10462-011-9272-4/METRICS.

Krippendorff, K. (2018) *Content Analysis: An Introduction to Its Methodology*. Thousand Oaks: Sage Publications.

Kumar, A., Gopal, R.D., Shankar, R. and Tan, K.H. (2022) Fraudulent review detection model focusing on emotional expressions and explicit aspects: investigating the potential of feature engineering. *Decision Support Systems*. 113728.

Kuratko, D.F. (2016) *Entrepreneurship: Theory, process, and practice*. Cengage learning.

Kwon, H.-J., Ban, H.-J., Jun, J.-K. and Kim, H.-S. (2021) Topic modeling and sentiment analysis of online review for airlines. *Information*. 12 (2), 78.

Kwon, O., Lim, S. and Lee, D.H. (2018) Acquiring startups in the energy sector: a study of firm value and environmental policy. *Business Strategy and the Environment*. 27 (8), 1376–1384. doi:10.1002/bse.2187.

Lai, X., Wang, F. and Wang, X. (2021) Asymmetric relationship between customer sentiment and online hotel ratings: the moderating effects of review characteristics. *International Journal of Contemporary Hospitality Management*. 33 (6), 2137–2156. doi:10.1108/IJCHM-07-2020-0708.

Landauer, T.K., Foltz, P.W. and Laham, D. (1998) An introduction to latent semantic analysis. *Discourse Processes*. 25 (2–3), 259–284.

Lebret, R. and Collobert, R. (2014) *N-gram-Based Low-Dimensional Representation for Document Classification*. http://arxiv.org/abs/1412.6277.

Lee, H., Kelley, D., Lee, J. and Lee, S. (2012) SME Survival: The Impact of Internationalization, Technology Resources, and Alliances. *Journal of Small Business Management*. 50 (1), 1–19. doi:10.1111/j.1540-627X.2011.00341.x.

Leeman, D. (2007) *Topic Modeling with Latent Dirichlet Allocation*.

Lewis, D.D., Yang, Y., Russell-Rose, T. and Li, F. (2004) Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*. 5 (Apr), 361–397.

Li, L., Ma, S., Han, X., Zheng, C. and Wang, D., 2021. Data-driven online service supply chain: a demand-side and supply-side perspective. *Journal of Enterprise Information Management*, *34*(1), pp.365-381.

Li, X., Wu, C. and Mai, F. (2019) The effect of online reviews on product sales: A joint sentiment-topic analysis. *Information and Management*. 56 (2), 172–184.

Liew, W. Te, Adhitya, A. and Srinivasan, R. (2014) Sustainability trends in the process industries: A text mining-based analysis. *Computers in Industry*. 65 (3), 393–400. doi:https://doi.org/10.1016/j.compind.2014.01.004.

Lin, F.R., Hsieh, L.S. and Chuang, F.T. (2009) Discovering genres of online discussion threads via text mining. *Computers and Education*. 52 (2), 481–495. doi:10.1016/J.COMPEDU.2008.10.005.

Ling, T. and Dobbie, G. (2004) *Semistructured database design*. Springer Science and Business Media.

Liu, B. (2012) Sentiment analysis and opinion mining. In: *Synthesis lectures on human language technologies*. pp. 1–167. doi:10.2200/S00416ED1V01Y201204HLT016).

Liu, B. (2010) Sentiment analysis and subjectivity. In: *Handbook of natural language processing*. pp. 627–666.

Liu, X., Lee, D. and Srinivasan, K. (2019) Large-scale cross-category analysis of consumer review content on sales conversion leveraging deep learning. *Journal of Marketing Research*. 56 (6), 918–943.

Lo, A.S. and Yao, S.S. (2019) What makes hotel online reviews credible? *International Journal of Contemporary Hospitality Management*. 31 (1), 41–60. doi:10.1108/IJCHM-10-2017-0671.

Loock, M. (2012) Going beyond best technology and lowest price: on renewable energy investors' preference for service-driven business models. *Energy Policy*. 40, 21–27.

Losee, R.M. (2006) Browsing mixed structured and unstructured data. *Information Processing and Management*. 42 (2), 440–452. doi:10.1016/j.ipm.2005.02.001.

Lu, S., Xiao, L. and Ding, M. (2016) A Video-Based Automated Recommender (VAR) System for Garments. *Marketing Science*. 35 (3), 484–510. doi:10.1287/mksc.2016.0984.

Lu, W. and Stepchenkova, S. (2015) Journal of Hospitality Marketing and Management User-Generated Content as a Research Mode in Tourism and Hospitality Applications: Topics, Methods, and Software. *Journal of Hospitality Marketing and Management*. 24 (2), 119–154. doi:10.1080/19368623.2014.907758.

Lundberg, S.M. and Lee, S.-I. (2017) A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017 pp. 4768–4777.

Mai, L. and Le, B. (2021) Joint sentence and aspect-level sentiment analysis of product comments. *Annals of Operations Research*. 300 (2), 493–513. doi:10.1007/s10479-020-03534-7.

De Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J. and Manning, C.D. (2014) Universal Stanford dependencies: A cross-linguistic typology. In: *LREC*. 2014 pp. 4585–4592.

Marra, A., Antonelli, P. and Pozzi, C. (2017) Emerging green-tech specializations and clusters – A network analysis on technological innovation at the metropolitan level. *Renewable and Sustainable Energy Reviews*. 67, 1037–1046. doi:10.1016/J.RSER.2016.09.086.

Marra, A., Carlei, V. and Baldassari, C. (2020) Exploring networks of proximity for partner selection, firms' collaboration and knowledge exchange. The case of clean-tech industry. *Business Strategy and the Environment*. 29 (3), 1034–1044. doi:10.1002/bse.2415.

Marshan, A., Kansouzidou, G. and Ioannou, A. (2020) Sentiment Analysis to Support Marketing Decision Making Process: A Hybrid Model. In: *Proceedings of the Future Technologies Conference*. Springer. pp. 614–626.

Martin, R.E. and Justis, R.T. (1993) Franchising, liquidity constraints and entry. *Applied Economics*. 25 (9), 1269–1277.

Masini, A. and Menichetti, E. (2012) The impact of behavioural factors in the renewable energy investment decision making process: Conceptual framework and empirical findings. *Energy Policy*. 40, 28–38.

May, T. and Perry, B. (2022) *Social research: Issues, methods and process*. McGraw-Hill Education (UK).

Mazzucato, M. and Semieniuk, G. (2018) Financing renewable energy: Who is financing what and why it matters. *Technological Forecasting and Social Change*. 127, 8–22. doi:10.1016/J.TECHFORE.2017.05.021.

Mihalcea, R. and Strapparava, C. (2009) The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language. In *Proceedings of the ACL-IJCNLP 2009 conference*. 309–312.

Mihalcea, R. and Tarau, P., 2004, July. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing* (pp. 404-411).

Mimno, D., Wallach, H., Talley, E., Leenders, M. and McCallum, A. (2011) Optimizing semantic coherence in topic models. In: *Proceedings of the 2011 conference on empirical methods in natural language processing*. pp. 262–272.

Minka, T.P. and Lafferty, J. (2002) Expectation-propagation for the generative aspect model. In: *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*. pp. 352–359.

Mochkabadi, K. and Volkmann, C.K. (2020) Equity crowdfunding: a systematic review of the literature. *Small Business Economics*. 54 (1), 75–118. doi:10.1007/s11187-018-0081-x.

Molnar, C. (2020) *Interpretable machine learning: A guide for making Black Box Models interpretable.* Lulu.

Moreno-Perdigón, M.C., Guzmán-Pérez, B. and Mesa, T.R. (2021) Guest satisfaction in independent and affiliated to chain hotels. *International Journal of Hospitality Management*. 94, 102812.

Mrkajic, B., Murtinu, S. and Scalera, V.G. (2019) Is green the new gold? Venture capital and green entrepreneurship. *Small Business Economics*. 52 (4), 929–950. doi:10.1007/s11187-017-9943-x.

Müller, O., Junglas, I., Debortoli, S. and Vom Brocke, J. (2016) Using text analytics to derive customer service management benefits from unstructured data. . *MIS Quarterly Executive*. 15 (4).

Nadeau, D. and Sekine, S. (2007) A survey of named entity recognition and classification. *Lingvisticae Investigationes*. 30 (1), 3–26. doi:10.1075/LI.30.1.03NAD.

Namkung, Y. and Jang, S. (2007) Does Food Quality Really Matter in Restaurants? Its Impact On Customer Satisfaction and Behavioral Intentions. *Journal of Hospitality and Tourism Research*. 31 (3), 387–409. doi:10.1177/1096348007299924.

Nasr, L., Burton, J. and Gruber, T. (2018) Developing a deeper understanding of positive customer feedback. *Journal of Services Marketing*. 32 (2), 142–160. doi:10.1108/JSM-07-2016-0263/FULL/PDF.

Neligan, A., Baumgartner, R.J., Geissdoerfer, M. and Schöggl, J. (2023) Circular disruption: Digitalisation as a driver of circular economy business models. *Business Strategy and the Environment*. 32 (3), 1175–1188. doi:10.1002/bse.3100.

Nijmeijer, K.J., Fabbricotti, I.N. and Huijsman, R. (2014) Making franchising work: A framework based on a systematic review. *International Journal of Management Reviews*. 16 (1), 62–83.

Nisbet, R., Elder, J. and nd Miner, G.D. (2009) *Handbook of statistical analysis and data mining applications*. Academic press.

Nivre, J. (2010) Dependency Parsing. *Language and Linguistics Compass*. 4 (3), 138–152. doi:10.1111/j.1749-818X.2010.00187.x.

Norton, S.W. (1988) An Empirical Look at Franchising as an Organizational Form. *The Journal of Business*. 61 (2), 197–218. http://www.jstor.org/stable/2352900.

Pang, B., Lee, L. and Vaithyanathan, S. (2002) Thumbs up? Sentiment classification using machine learning techniques. In: *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*. p. doi:10.3115/1118693.1118704.

Pantano, E., Dennis, C. and Alamanos, E. (2021) Retail managers' preparedness to capture customers' emotions: a new synergistic framework to exploit

unstructured data with new analytics. *British Journal of Management*. 33(3), 1179-1199.

Parasuraman, A., Zeithaml, V.A. and Berry, L.L. (1985) A conceptual model of service quality and its implications for future research. *Journal of Marketing*. 49 (4), 41–50.

Park, D.-H., Lee, J. and Han, I. (2007) The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement. *International Journal of Electronic Commerce*. 11 (4), 125–148.

Pavlou, P.A. and Dimoka, A. (2006) The nature and role of feedback text comments in online marketplaces: Implications for trust building, price premiums, and seller differentiation. *Information Systems Research*. 17 (4), 392–414.

Payne, A.F., Storbacka, K. and Frow, P. (2008) Managing the co-creation of value. *Journal of the Academy of Marketing Science*. 36 (1), 83–96. doi:10.1007/s11747-007-0070-0.

Pecina, P., 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation*. *44*, 137-158.

Perryman, A.A. and Combs, J.G. (2012) Who should own it? An agency-based explanation for multi-outlet ownership and co-location in plural form franchising. *Strategic Management Journal*. 33 (4), 368–386.

Pian, W., Khoo, C. and Chang, Y. (2014) Relevance Judgment When Browsing a Health Discussion Forum: Content Analysis of Eye Fixations. *Library and Information Science Research Electronic Journal*. 24 (2), 132–147. https://www.libres-ejournal.info/wp-content/uploads/2015/03/LIBRESv24i2p132-147.Pian_.2014.pdf.

Pinelli, F., Davachi, L. and Higgins, E.T. (2022) Shared Reality Effects of Tuning Messages to Multiple Audiences. *Social Cognition*. 40 (2), 172–183.

Pitelis, A., Vasilakos, N. and Chalvatzis, K. (2020) Fostering innovation in renewable energy technologies: Choice of policy instruments and effectiveness. *Renewable Energy*. 151, 1163–1172.

Piva, E. and Rossi-Lamastra, C. (2018) Human capital signals and entrepreneurs' success in equity crowdfunding. *Small Business Economics*. 51, 667–686.

Plummer, L.A., Allison, T.H. and Connelly, B.L. (2016) Better together? Signaling interactions in new venture pursuit of initial external capital. *Academy of Management Journal*. 59 (5), 1585–1604.

Polas, M.R.H., Rahman, M.M., Miah, M.A. and Hayash, M.M.A. (2018) The impact of waiting time towards customers satisfaction in fast food establishments: Evidence from Bangladesh. *IOSR Journal of Business and Management*. 20 (5), 11–21.

Pollock, T.G., Chen, G., Jackson, E.M. and Hambrick, D.C. (2010) How much prestige is enough? Assessing the value of multiple types of high-status affiliates for young firms. *Journal of Business Venturing*. 25 (1), 6–23. doi:10.1016/J.JBUSVENT.2009.01.003.

Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P. and Welling, M. (2008) Fast collapsed gibbs sampling for latent dirichlet allocation. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. pp. 569–577.

Porter, M.E. and Strategy, C. (1980) Techniques for analyzing industries and competitors. *Competitive Strategy.* New York: Free.

Prasad, A.K. (2000) Stereoscopic particle image velocimetry. *Experiments in Fluids*. 29 (2), 103–116.

Puri, M. and Zarutskie, R. (2012) On the life cycle dynamics of venture-capital-and non-venture-capital-financed firms. *The Journal of Finance*. 67 (6), 2247–2293.

Qiu, J., Liu, C., Li, Y. and Lin, Z. (2018) Leveraging sentiment analysis at the aspects level to predict ratings of reviews. *Information Sciences*. 451, 295–309. doi:10.1016/j.ins.2018.04.009.

Quan, C. and Ren, F. (2014) Unsupervised product feature extraction for feature-oriented opinion determination. *Information Sciences*. 272, 16–28.

Radhakrishna, V., Kumar, G.R. and Aljawarneh, S., 2017. Optimising business intelligence results through strategic application of software process model. *International Journal of Intelligent Enterprise*, *4*(1-2), 128-142.

Rahi, S. (2017) Research Design and Methods: A Systematic Review of Research Paradigms, Sampling Issues and Instruments Development. *International Journal of Economics and Management Sciences*. 6 (2), 1-5. doi:10.4172/2162-6359.1000403.

Rajendran, S. and Fennewald, J., 2021. Improving service supply chain of internet services by analyzing online customer reviews. *Supply Chain Management in Manufacturing and Service Systems: Advanced Analytics for Smarter Decisions*, pp.147-163.

Rao, S., Griffis, S.E. and Goldsby, T.J. (2011) Failure to deliver? Linking online order fulfillment glitches with future purchase behavior. *Journal of Operations Management*. 29 (7–8), 692–703.

Rao, V.C.S., Radhika, P., Polala, N. and Kiran, S. (2021) Logistic Regression versus XGBoost: Machine Learning for Counterfeit News Detection. In: *2021 Second International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)*. 2021 IEEE. pp. 1–6.

Reuer, J.J., Tong, T.W. and Wu, C.-W. (2012) A signaling theory of acquisition premiums: Evidence from IPO targets. *Academy of Management Journal*. 55 (3), 667–683.

Roberts, C.W. (2000) A Conceptual Framework for Quantitative Text Analysis On Joining Probabilities and Substantive Inferences about Texts. *Quality and Quantity*. 34, 259–274.

Roberts, M.E., Stewart, B.M. and Airoldi, E.M. (2016) A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*. 111 (515), 988–1003.

Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K., Albertson, B. and Rand, D.G. (2014) Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*. 58 (4), 1064–1082. doi:10.1111/ajps.12103.

Ruiz-Mafe, C., Chatzipanagiotou, K. and Curras-Perez, R. (2018) The role of emotions and conflicting online reviews on consumers' purchase intentions. *Journal of Business Research*. 89, 336–344.

Rychlý, P. (2018) A lexicographer-friendly association score . In: *Proceedings of recent advances in Slavonic natural language processing*. Brno: Masaryk University. pp. 6–9.

Ryu, K., Lee, H. and Kim, W.G. (2012) The influence of the quality of the physical environment, food, and service on restaurant image, customer perceived value, customer satisfaction, and behavioral intentions. *International Journal of Contemporary Hospitality Management*. 24, 200–223.

Sampson, S.E. and Spring, M. (2012) Customer roles in service supply chains and opportunities for innovation. *Journal of Supply Chain Management*. 48 (4), 30–50.

Sawant, R.J., Hada, M. and Blanchard, S.J. (2021) Contractual discrimination in franchise relationships. *Journal of Retailing*. 97 (3), 405–423.

Schönwälder, J. and Weber, A. (2023) Maturity levels of sustainable corporate entrepreneurship: The role of collaboration between a firm's corporate venture and corporate sustainability departments. *Business Strategy and the Environment*. 32 (2), 976–990. doi:10.1002/BSE.3085.

Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D. and Tufano, P., 2012. Analytics: The real-world use of big data. *IBM Global Business Services*, *12*(2012), pp.1-20.

Sebastiani, F. (2002) Machine Learning in Automated Text Categorization. *ACM Computing. Surveys.* 34 (1), 1–47. doi:10.1145/505282.505283.

See-To, E.W.K. and Ngai, E.W.T. (2018) Customer reviews for demand distribution and sales nowcasting: a big data approach. *Annals of Operations Research*. 270 (1), 415–431.

Shahin, A., 2010. SSCM: Service supply chain management. *International Journal of Logistics Systems and Management*, *6*(1), pp.60-75.

Shahzad, U., Ferraz, D., Nguyen, H.H. and Cui, L. (2022) Investigating the spill overs and connectedness between financial globalization, high-tech industries and environmental footprints: Fresh evidence in context of China. *Technological Forecasting and Social Change*. 174. doi:10.1016/J.TECHFORE.2021.121205.

Sharma, S.S. and Dutta, G. (2021) SentiDraw: Using star ratings of reviews to develop domain specific sentiment lexicon for polarity determination. *Information Processing and Management*. 58 (1), 102412.

Silva, R., Gerwe, O. and Becerra, M. (2017) Corporate brand and hotel performance: A resource-based perspective. *Journal of Business Research*. 79, 23–30.

Smith, J.A. (2017) Textual Analysis. *The International Encyclopedia of Communication Research Methods*. 1–7. doi:10.1002/9781118901731.IECRM0248.

Sorenson, O. and Sørensen, J.B. (2001) Finding the right mix: Franchising, organizational learning, and chain performance. *Strategic Management Journal*. 22 (6-7), 713–724.

Soucy, P. and Mineau, G.W. (2001) A simple KNN algorithm for text categorization. *Proceedings - IEEE International Conference on Data Mining, ICDM*. 647–648. doi:10.1109/ICDM.2001.989592.

Spence, M. (1978) Job Market Signaling. *Uncertainty in Economics*. 281–306. doi:10.1016/B978-0-12-214850-7.50025-5.

Standifird, S.S. (2001) Reputation and e-commerce: eBay auctions and the asymmetrical impact of positive and negative ratings. *Journal of Management*. 27 (3), 279–295.

Steigenberger, N. and Wilhelm, H. (2018) Extending Signaling Theory to Rhetorical Signals: Evidence from Crowdfunding. *Organization Science*. 29 (3). doi:10.1287/orsc.2017.1195.

Stern, N. and Valero, A. (2021) Innovation, growth and the transition to net-zero emissions. *Research Policy*. 50 (9), 104293. doi:10.1016/J.RESPOL.2021.104293.

Stevenson, R., McMahon, S.R., Letwin, C. and Ciuchta, M.P. (2022) Entrepreneur fund-seeking: toward a theory of funding fit in the era of equity crowdfunding. *Small Business Economics*. 58 (4), 2061–2086. doi:10.1007/s11187-021-00499-0.

Su, Y. and Teng, W. (2018). Contemplating museums' service failure: Extracting the service quality dimensions of museums from negative on-line reviews. *Tourism Management*, 69, 214-222.

Sulek, J.M. and Hensley, R.L. (2004) The relative importance of food, atmosphere, and fairness of wait: The case of a full-service restaurant. *Cornell Hotel and Restaurant Administration Quarterly*. 45 (3), 235–247.

Syed, S. and Spruit, M. (2017) Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In: *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. 2017 IEEE. pp. 165–174. doi:10.1109/DSAA.2017.61.

Symitsi, E., Stamolampros, P., Daskalakis, G. and Korfiatis, N. (2021) The informational value of employee online reviews. *European Journal of Operational Research*. 288 (2), 605–619.

Taddy, M. (2012) On estimation and selection for topic models. In: *Artificial Intelligence and Statistics*. 2012 PMLR. pp. 1184–1193.

Tan, Y., Zhang, M., Liu, Y. and Ma, S. (2016) Rating-boosted latent topics: Understanding users and items with ratings and reviews. In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. 2016 pp. 2640–2646.

Thakur, R. (2018) Customer engagement and online reviews. *Journal of Retailing and Consumer Services*. 41, 48–59. doi:10.1016/J.JRETCONSER.2017.11.002.

Thelwall, M., Wilkinson, D. and American, S.U. (2009) Data mining emotion in social network communication: Gender differences in MySpace. *Journal of the American Society for Information Science and Technology*. 61 (1), 190–199. doi:10.1002/asi.21180.

Tirunillai, S. and Tellis, G.J. (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. doi:10.1509/jmr.12.0106.

Trilling, D. and Jonkman, J.G.F. (2018) Scaling up Content Analysis. *Communication Methods and Measures*. 12 (2–3), 158–174. doi:10.1080/19312458.2018.1447655.

Truong, T.J., Ludwig, S., Mooi, E. and Bove, L. (2022) The market value of rhetorical signals in technology licensing contracts. *Industrial Marketing Management*. 105, 489–501.

Tseng, M.-L., Lim, M.K., Wong, W.-P., Chen, Y.-C. and Zhan, Y. (2018) A framework for evaluating the performance of sustainable service supply chain management under uncertainty. *International Journal of Production Economics*. 195, 359–372.

Tumasjan, A., Braun, R. and Stolz, B. (2021) Twitter sentiment as a weak signal in venture capital financing. *Journal of Business Venturing*. 36 (2), 106062. doi:10.1016/j.jbusvent.2020.106062.

Vanacker, T. and Forbes, D.P. (2016) Disentangling the multiple effects of affiliate reputation on resource attraction in new firms. *Organization Science*. 27 (6), 1525–1547.

Vayansky, I. and Kumar, S.A.P. (2020) A review of topic modeling methods. *Information Systems*. 94, 101582. doi:10.1016/J.IS.2020.101582.

Verma, S. and Yadav, N. (2021) Past, present, and future of electronic word of mouth (EWOM). *Journal of Interactive Marketing*. 53, 111–128.

Vijayarani, S., Ilamathi, M.J. and Nithya, M. (2015) Preprocessing techniques for text mining-an overview. *International Journal of Computer Science and Communication Networks*. 5 (1), 7–16.

Vismara, S. (2016) Equity retention and social network theory in equity crowdfunding. *Small Business Economics*. 46 (4), 579–590. doi:10.1007/s11187-016-9710-4.

Wachsmuth, H. (2015) *Text analysis pipelines towards ad-hoc large-scale text mining*. Weimar: Springer Cham.

Wang, Y. and Chaudhry, A. (2018) When and how managers' responses to online reviews affect subsequent reviews. *Journal of Marketing Research*. 55 (2), 163–177.

Wang, Y., Li, X., Zhang, L.L. and Mo, D. (2022) Configuring products with natural language: a simple yet effective approach based on text embeddings and multilayer perceptron. *International Journal of Production Research*. 60 (17), 5394–5406. doi:10.1080/00207543.2021.1957508.

Wang, Y., Wallace, S.W., Shen, B. and Choi, T.-M. (2015) Service supply chain management: A review of operational models. *European Journal of Operational Research*. 247 (3), 685–698. doi:https://doi.org/10.1016/j.ejor.2015.05.053.

Wang, Y., Zhong, K. and Liu, Q. (2022) Let criticism take precedence: Effect of side order on consumer attitudes toward a two-sided online review. *Journal of Business Research*. 140, 403–419.

Wehnert, P. and Beckmann, M. (2021) Crowdfunding for a sustainable future: A systematic literature review. *IEEE Transactions on Engineering Management*.

Wei, P.-S. and Lu, H.-P. (2013) An examination of the celebrity endorsements and online customer reviews influence female consumers' shopping behavior. *Computers in Human Behavior*. 29 (1), 193–201.

Wei, Y., Hu, Q. and Xu, C. (2013) Ordering, pricing and allocation in a service supply chain. *International Journal of Production Economics*. 144 (2), 590–598.

Wiklund, J. and Shepherd, D. (2003) Knowledge-based resources, entrepreneurial orientation, and the performance of small and medium-sized businesses. *Strategic Management Journal*. 24 (13), 1307–1314.

Wirth, R. and Hipp, J. (2000) CRISP-DM: Towards a standard process model for data mining. In: *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. Manchester. pp. 29–39.

Wöhler, J. and Haase, E. (2022) Exploring investment processes between traditional venture capital investors and sustainable start-ups. *Journal of Cleaner Production*. 377, 134318. doi:10.1016/J.JCLEPRO.2022.134318.

Wu, J., Li, Y. and Ma, Y. (2021) Comparison of XGBoost and the Neural Network model on the class-balanced datasets. In: *2021 IEEE 3rd International Conference on Frontiers Technology of Information and Computer (ICFTIC)*. 2021 IEEE. pp. 457–461.

Wu, T.K., Huang, S.C. and Meng, Y.R. (2008) Evaluation of ANN and SVM classifiers as predictors to the diagnosis of students with learning disabilities. *Expert Systems with Applications*. 34 (3), 1846–1856. doi:10.1016/J.ESWA.2007.02.026.

Wu, Y., Ngai, E.W.T., Wu, P. and Wu, C. (2020) Fake online reviews: Literature review, synthesis, and directions for future research. *Decision Support Systems*. 132, 113280. doi:10.1016/J.DSS.2020.113280.

Xing, S., Wang, Q., Zhao, X. and Li, T. (2019) A hierarchical attention model for rating prediction by leveraging user and product reviews. *Neurocomputing*. 332, 417–427.

Xu, S. (2018) Bayesian Naïve Bayes classifiers to text classification. *Journal of Information Science*. 44 (1), 48–59. doi:10.1177/0165551516677946/ASSET/IMAGES/LARGE/10.1177_0165551516677946-FIG5.JPEG.

Xu, X. (2020) Examining an asymmetric effect between online customer reviews emphasis and overall satisfaction determinants. *Journal of Business Research*. 106, 196–210.

Xu, X. (2021) What are customers commenting on, and how is their satisfaction affected? Examining online reviews in the on-demand food service context. *Decision Support Systems*. 142, 113467.

Yadav, A. and Vishwakarma, D.K. (2020) Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*. 53 (6), 4335–4385.

Yan, Z., Wang, J., Dong, Q., Zhu, L., Lin, W. and Jiang, X. (2022) XGBoost algorithm and logistic regression to predict the postoperative 5-year outcome in patients with glioma. *Annals of translational medicine*. 10 (16), 860.

Yang, B., Liu, Y., Liang, Y. and Tang, M. (2019) Exploiting user experience from online customer reviews for product design. *International Journal of Information Management*. 46, 173–186. doi:https://doi.org/10.1016/j.ijinfomgt.2018.12.006.

Yang, F.J. (2018) An implementation of naive bayes classifier. *Proceedings - 2018 International Conference on Computational Science and Computational Intelligence, CSCI 2018*. 301–306. doi:10.1109/CSCI46756.2018.00065.

Ye, F., Xia, Q., Zhang, M., Zhan, Y. and Li, Y. (2022) *Harvesting Online Reviews to Identify the Competitor Set in a Service Business: Evidence From the Hotel Industry*. 25 (2), 301–327. doi:10.1177/1094670520975143.

Yeo, S.F., Tan, C.L., Kumar, A., Tan, K.H. and Wong, J.K. (2022) Investigating the impact of AI-powered technologies on Instagrammers' purchase decisions in digitalization era–A study of the fashion and apparel industry. *Technological Forecasting and Social Change*. 177, 121551.

Yu, C.H., Jannasch-Pennell, A. and DiGangi, S. (2011) Compatibility between Text Mining and Qualitative Research in the Perspectives of Grounded Theory, Content Analysis, and Reliability. *Qualitative Report*. 16 (3), 730–744. http://www.nova.edu/ssss/QR/QR16-3/yu.pdf.

Zelikovitz, S. and Hirsh, H. (2001) Using LSI for text classification in the presence of background text. In: *Proceedings of the tenth international conference on Information and knowledge management*. 2001 pp. 113–118.

Zhang, C., Tian, Y.-X. and Fan, L.-W. (2020) Improving the Bass model's predictive power through online reviews, search traffic and macroeconomic data. *Annals of Operations Research*. 295 (2), 881–922.

Zhang, S., Li, X., Zong, M., Zhu, X. and Cheng, D. (2017) Learning k for kNN Classification. *ACM Transactions on Intelligent Systems and Technology (TIST)*. 8 (3). doi:10.1145/2990508.

Zhang, W. and Wang, J. (2016) Integrating topic and latent factors for scalable personalized review-based rating prediction. *IEEE transactions on knowledge and data engineering*. 28 (11), 3013–3027.

Zhao, Y., Xu, X. and Wang, M. (2019) Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews. *International Journal of Hospitality Management*. 76, 111–121.

Zhao, Z., Wang, J., Sun, H., Liu, Y., Fan, Z. and Xuan, F. (2019) What factors influence online product sales? Online reviews, review system curation, online promotional marketing and seller guarantees analysis. *IEEE Access*. 8, 3920–3931.

Zhou, Y., Wang, X. and Yuen, K.F. (2021) Sustainability disclosure for container shipping: A text-mining approach. *Transport Policy*. 110, 465–477. doi:https://doi.org/10.1016/j.tranpol.2021.06.020.

Zhu, F. and Zhang, X. (2010) Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing*. 74 (2), 133–148.

Žižka, J., Dařena, F. and Svoboda, A. (2019) Text mining with Machine Learning: Principles and techniques. Boca Raton: CRC Press.

Zohuri, B., Mossavar-Rahmani, F. and Behgounia, F. (2022) Structured and unstructured data processing. In: *Knowledge is Power in Four Dimensions: Models to Forecast Future Paradigm*. Academic Press. pp. 87–119. doi:10.1016/B978-0-323-95112-8.00004-0.

Žukauskas, P., Vveinhardt, J. and Andriukaitienė, R. (2018) Philosophy and Paradigm of Scientific Research. In: *Management Culture and Corporate Social Responsibility*. IntechOpen. doi:10.5772/intechopen.70628.

Appendix

Table A1 Word clouds for 15 topics

 (1) Food quality (critique)	 (2) Repeat purchase	 (3) Customer rumination	 (4) Order issues (incorrectness)
 (5) Food quality (critique)	 (6) Delivery service (praise)	 (7) Delivery service (critique)	 (8) Value for money
 (9) Order issues (incompleteness)	 (10) Takeaway experience	 (11) Packaging issues	 (12) Meal deal
 (13) Packaging issues	 (14) Food quality (Praise)	 (15) Delivery time	